11-522
1988
c.3

Statistics Canada Statistique Canada

Ottawa, Canada

SYMPOSIUM 88

The Impact of High Technology on Survey Taking

October 24-25, 1988

Ottawa, Ontario

PROCEEDINGS

STATISTICS CANADA STATISTIQUE CANADA
MAY MAI 1 3 2008
LIBRARY
BIBLIOTHEQUE

Canada

SYMPOSIUM 88

The Impact of High Technology on Survey Taking

October 24-25, 1988

Ottawa, Ontario

PROCEEDINGS

Issued April, 1989

Compiled and edited by

J. Kovar and E. Doucet

PREFACE

Symposium 88 was the fifth international symposium on methodology and informatics issues sponsored by Statistics Canada. As in previous years, the objective of the symposium was to provide a forum for experts from around the world to exchange ideas of common interest.

The conference was held in the Simon Goldberg Conference Centre at Statistics Canada. More than 400 persons, representing over 20 countries, participated. The opening remarks were followed by Bo Sundgren's keynote address on "The Impact of New Technology on Methodology and Organization of Statistical Data Processing". Seventeen invited papers were presented in the remainder of the two day schedule. Emphasis was placed on innovative uses of high technology relating in particular to data collection, processing, analysis and dissemination.

This volume of proceedings contains a record of the papers as they were submitted. The views expressed therein are those of the authors and do not necessarily represent official views or policies of the organizations they work for, nor those of Statistics Canada. To facilitate the use of these proceedings, the contents of this volume follow the order of presentation in the programme.

The sixth international Statistics Canada Symposium will be held in Ottawa, in October 1989. Its focus will be the Analysis of Data in Time, with particular emphasis placed on the design and analysis of repeated surveys. We look forward to seeing you there and would like to take this opportunity to thank all those who have helped make Symposium 88 a success.

J.E.Doucet
Informatics Branch
Statistics Canada

J.G.Kovar
Methodology Branch
Statistics Canada

CONTENTS

Opening Remarks: I.P.Fellegi, Chief Statistician of Canada 1
(Delivered by **G.Brackstone,** Statistics Canada)

Keynote Address: The impact of new technology on methodology and organization of statistical data processing 7
B.Sundgren, Statistics Sweden

Session 1: Recent Developments
Chaired by **E.Outrata**, Statistics Canada

The use of new technologies at Statistics Canada 27
J.E.Doucet, Statistics Canada

The use of technology at Statistics Sweden 37
L.Lyberg and P.Dean, Statistics Sweden

HyperQuest: A new method for interviewing using laptop computers ... 63
M.A.Jaro, Systems Automation Corporation

Session 2: Data Collection
Chaired by **B.Petrie,** Statistics Canada

The impact of high technology on data collection 75
W.L.Nicholls II, US Bureau of the Census

The effects of CATI on cost and data quality 103
G.Catlin, S.Ingram and **L.Hunter,** Statistics Canada

Two, new, experimental data collection methods at INSEE 123
C.Bernard, Institut national de la statistique et des études économiques

Computer-based question banks: Why and how? 139
R.P.Baker, National Opinion Research Center

Data collection and capture for surveys 155
I.S.Green and M.J.Wade, Statistics Canada

Session 3: Data Processing
Chaired by **J.Ryten,** Statistics Canada

Computer assisted survey processing on PC and PC-LAN 171
W.J.Keller and J.G.Bethlehem,
Netherlands Central Bureau of Statistics

The use of microcomputers for census processing in developing countries: An update 181
V.Toro and K.Chamberlain, US Bureau of the Census

A report on recent survey processing in developing countries: The Demographic and Health Surveys microcomputer approach 201
J.Cushing, Institute for Resource Development/Westinghouse

A comparison of software for editing survey and census data 211
P.Cotton, Statistical Consultant

Session 4: Data Analysis
Chaired by **J.N.K.Rao,** Carleton University

Software for survey analysis 243
W.A.Fuller, Iowa State University

SUDAAN: A comprehensive software package for survey data analysis .. 257
L.M.LaVange, **B.V.Shah,** B.G.Barnwell and J.F.Killinger,
Research Triangle Institute

Session 5: A Future Outlook
Chaired by **G.Brackstone,** Statistics Canada

Contributions of AI to statistical data processing 277
J.H.Johnson, Statistics Canada

A technological forecast of the role of personal computers in statistical survey methods 295
J.C.Nash, University of Ottawa

OPENING REMARKS

I. P. FELLEGI,

Chief Statistician of Canada

Delivered by G. Brackstone,
Statistics Canada

WELCOME

Good morning; and a very warm welcome to Symposium 88. This is the fifth in this series of symposia, sponsored by Statistics Canada and the Laboratory for Research in Statistics and Probability, Carleton University and the University of Ottawa.

The program which the organizers have prepared for us this year is an impressive one. I was especially pleased to note the good level of international participation - with strong representation from:

- Statistics Sweden,
- L' institut national de la statistique et des etudes economiques, in France,
- the Netherlands Bureau of Statistics, and
- the U.S. Bureau of the Census;

as well as strong representation from several prominent universities and private corporations, in Canada and in the U.S.

From Statistics Canada and the city of Ottawa, may I offer you a warm WELCOME et BIENVENUE!

Welcome too, to our large contingent from Statistics Canada. I see that we are well represented: both in the formal program, as well as in numbers among the audience.

COMMENTS ON THE CHOICE OF TOPIC

Previous symposia in this series have focused primarily on topics directly related to survey methodology. Over the years, we have we have met in similar situations to study the following issues:

- Analysis of Survey Data (1984);
- Small Area Statistics (1985);
- Missing Data in Surveys (1986); and
- Statistical Uses of Administrative Data (1987).

This year's topic, THE IMPACT OF HIGH TECHNOLOGY ON SURVEY TAKING, represents a mild departure from the trend established over the past four years. The topic is clearly positioned at the intersection of survey methodology and informatics; and I find this to be highly appropriate at this time. Why? Because the topic poses a timely challenge; both, to the SURVEY METHODOLOGIST, and to the INFORMATICS SPECIALIST. Permit me to speculate for a few moments on the nature of that challenge.

The Challenge to Informatics

Let us look first at the nature of the challenge to the Informatics Specialist. In recent years, Statistics Canada has experienced a period of unprecedented growth and innovation in computing technology. The examples are all around us.

- We now have some 1200 microcomputer-based workstations in this organization; four years ago, we had none.
- We have a modern mainframe facility at Statistics Canada whose capacity has been growing at roughly the rate of 30% annually over the past ten years.
- We have the latest relational database software installed across different operating system platforms, permitting very rapid development of quite sophisticated applications.
- We have also begun to exploit technologies which, until quite recently, were regarded as esoteric: interactive graphics workstations, optical disk storage systems, desk-top publishing systems, etc.

Yet, although the tangible signs of progress are all around us, many of the problems with which we have wrestled for some fifteen years, persist even today. Consider just a few these persistent problems.

- PROBLEM - An effective and affordable automated method for dealing with Disclosure Avoidance still manages to elude us.
- PROBLEM - Survey Design and Specification is still inadequately aided by the awesome computing power available to us. And many analysts remain all too willing to construct yet-another editing system, for yet-another survey - content to re-invent the wheel and to make little adjustments on the margin.
- PROBLEM - Despite the availability of very powerful proprietary software, Data Retrieval and Assembly is still seriously constrained by the difficulty of navigating the maze of technologies across which the target data are stored.

These are just a few of these persistent problems; you probably have your own lists. Symposium 88 challenges the Informatics Specialist to explore 'de nouveau' the possibilities for applying the fruits of all this technological innovation to more directly benefit the survey taking process.

The Challenge to Methodology

For the Survey Methodologist the challenge posed by this symposium is perhaps less obvious.

The availability of relatively cheap and powerful computing resources has opened the door to re-evaluating our choice of survey taking methods and techniques. Permit me to illustrate what I mean, with a few examples:

- Computer-intensive statistical techniques have become increasingly feasible (in the economic and technical sense) and there exist possibilities for applying these techniques in many interesting ways: multiple stage sampling, simulation and modelling, complex data analysis, exhaustive approaches to record linkage, to name just a few. We must routinely question whether our approaches to problem solving are appropriate in light of these new possibilities; or whether we are professionally bowing to "standard practice" and "conventional wisdom" because we are experienced (yes, even comfortable) with the existing techniques.

- A database-driven approach to survey processing has become genuinely feasible, with the advent of commercially viable relational database management system software. The database-driven approach is significant in that it has the potential to break down the traditional 'time-sequential-batch' model of survey processing - a model which reflects the technological constraints imposed by the earlier stages of computerization. Our existing model of survey processing looks like it does, because the model had to be compromised to work on the earlier generation of computers. With the availability of relational database processors, it may now be time to 'rethink' our survey processing model.

Symposium 88 challenges the Survey Methodologist to examine the current repertoire of models, tools and techniques; and to assess its appropriateness today - in view of the expanded range of possibilities offered by the newer technologies. Even more important, we need to see whether transforming our problems in non-traditional ways might lead to the solutions that have escaped us to date.

Synergy

Of course, the challenge to Informatics and the challenge to Methodology cannot be as easily compartmentalized in reality, as they can in an opening address of this nature. Increasingly, the reality is that:

- the effective survey methodologist is very computer literate; and

- the effective Informatics Specialist has a good understanding of statistics and survey methodology.

That is why we can meet together, in the manner of Symposium 88, to address these new challenges in common.

And, this brings us back to the intersection of Informatics and Methodology - and the synergy which mechanisms like Symposium 88 can bring to advancing the state of our art. The theme of our meeting is well chosen. It is, and will remain, timely - for years to come.

INTRODUCTION OF KEYNOTE SPEAKER

I now turn to the pleasant task of introducing the keynote speaker for Symposium 88 - who is our colleague (Dr.) Bo Sundgren from Sweden.

Bo Sundgren is -

- Head of the EDP Systems Department - Statistics Sweden;
- Associate Professor - University of Stockholm; and
- Associate Professor - Stockholm School of Economics.

Statistics Canada's association with Bo Sundgren has been a long and profitable one. Those of you whose experience in statistical data processing dates back to the early seventies will remember Bo Sundgren very well. He was an important pioneer in the related fields of Data Modelling and Data Base Design - a pioneer who was well acquainted with the special requirements of a statistical office. In particular, he was a key player in the early efforts to develop a relational processor for Census systems - at a time when none was available in the commercial marketplace. His efforts contributed to the development of the RAPID system, which is currently in use in many countries (including this one) today.

Our association with Bo Sundgren has continued over the years, primarily in the context of his work with the Statistical Computing Program of the United Nations and the Conference of European Statisticians. In that area of activity, he is generally acknowledged to be the guiding-light in the international effort to arrive at a common suite of survey processing software, which could be applied across the major national statistical offices - and which could be used as the basis of a technology transfer to less developed nations.

Bo Sundgren is an important thinker, who has never been reticent to examine the fundamental assumptions which often underly our approaches to problem solving. I know that he will stimulate your thought processes - and set the tone for what I hope will be a rewarding and enjoyable symposium.

Dr. Sundgren, welcome to Symposium 88.

KEYNOTE ADDRESS: THE IMPACT OF NEW TECHNOLOGY ON METHODOLOGY AND ORGANIZATION OF STATISTICAL DATA PROCESSING

B. SUNDGREN,

Statistics Sweden

THE IMPACT OF NEW TECHNOLOGY ON METHODOLOGY AND ORGANIZATION OF STATISTICAL DATA PROCESSING

Abstract. During the last few decades we have witnessed a fantastic technological development, and an unbelievable reduction of the price/performance ratio for computers. Statistical offices have benefitted a lot from this development, by rationalizing their survey operations. We can now expect the new technology to be used more systematically in tasks requiring more "intelligence" than the basic computations and data handling operations in statistics production. Statistical design and EDP systems design are examples of such tasks. The new technology is also bringing about qualitative and structural changes, both inside statistical offices, and in the relations between statistical offices and its customers. Some effects, which are already becoming visible, are (i) standardization of technology, software, and methodology; (ii) decentralization of computers and computer-related resources; and (iii) integration of different types of tasks, skills, and competencies. In-house software development is being critically examined by several statistical offices, and it is an open question whether we shall see new generations of generalized software products specially destined for statistics production, and what they will look like. International cooperation in software development between statistical offices could be an attractive alternative. Outside statistical offices the new technology will offer new possibilities to the users of statistics, and the users will demand better service from statistics producers. This may call for some rethinking and reorganization within statistical offices.

1 The historical development and today's situation

The impressiveness of the technological development and the capacity and inexpensiveness of today's computers is well-known and need not be repeated here. It may be enough to point to the fact that each one of the personal computers standing today on the desks of individual staff members of a statistical office has approximately the same capacity as the whole mainframe of the statistical office 20 years ago, but at a fraction of its price. I think that it is also fair to say that most of these personal computers are not just standing there on the desks as a kind of status symbol, but they are being used very productively in the work of the statistical office.

It may be interesting to contrast this actual situation with some forecasts about the needs for computers that were made a few decades ago, when this technology emerged and became practically available. At that time some of the leading computer experts in Sweden seriously believed that one computer would be more than enough for all the needs of our country for the foreseeable future. Similar judgments were made in other countries. This type of prognoses is all the more remarkable, since if we look at the **functionality** (rather than capacity and price) of computers, the development has not been that dramatical. On a low, technical level, computers do essentially the same things now as then, only so enormously much more efficiently. Thus in principle the computer experts of 30 - 40 years ago should have been able to prognosticate a little more accurately the potential of computer technology. But they did not. The mistake they made was that they considered only one narrow category of applications, mathematical computations, and even for that type of application they did not have the imagination to foresee the explosion of needs that would appear, once the technology was available on a large scale and at a minimal price.

One thing I want to say with this is that from the very start of computer history, we seem to have been lagging behind in our ability to fully appreciate the application potentials of computer technology, and to actively plan for the most constructive usage of these potentials. This seems to be true also for statistical offices. We are eager to acquire the most recent computer technology, but, in my opinion, we are far too often spending too much of our resources just to move the same old applications between different generations of technologies, rather than actively developing new applications, new methodologies, and new ways of performing the overall tasks of a statistical office, based on a little more imaginative, long-term, strategical judgments about the future availability of computer technology. We also seem to neglect the drastic changes in cost relationships that are taking place all the time between hardware, software, and personnel resources, and to fail to ask ourselves more explicitly now and then whether not quite new mixtures of these production factors would be more optimal. For example, at Statistics Sweden, until recently, the acquirement of a personal computer had to be formally approved by the Director General of our agency, whereas the employment of a secretary, an investment that is roughly 1000 times bigger than the purchase of a PC, and with a 10 times longer "write-off period", could be decided on a lower managerial level.

2 Different types and levels of computer support

In a statistical agency there is a wide spectrum of possibilities to use computers. Some of these possibilities have already been exploited to a great extent, whereas others are at best in the prospecting stage. In order to discuss the potential of modern technology in statistical work, we need a basic structuring of that work. I have chosen to use two alternative structures. One is a classification of statistical tasks into those which are of a more or less routine character, and those which seem to require more "intelligence" of one sort or another. According to the other classification we make a distinction between the statistical operations as such, on the one hand, and the control of statistical operations on the other, where "control" includes planning, administration, and evaluation. Among the control tasks we may again distinguish between those which aim at individual statistical surveys, and those which have a whole statistical system as their object, for example, the whole statistical system of a country.

2.1 Routine applications vs "intelligence"

The vast majority of computer applications in a statistical office today are of a rather routine nature. Data are entered, edited, sorted, counted, and presented in a fairly straight-forward way. The computations are not always very complicated, but the volumes of data are sometimes quite large. The computer is little more than a pedantic, incredibly efficient book-keeper, who makes no errors. Nevertheless, this has turned out to be good enough to save large amounts of money for statistical agencies.

However, the challenge that is now facing us is whether we can start using computers in a more "intelligent" way. So far we have been very successful in multiplying the human being's capability to move and sort data, and to count them, and to eliminate the human tendency to commit errors in those operations. But can we also use the computer as an amplifier of the human intellect in statistical work? Without exaggerating the possibilities of disciplines with fancy names like "artificial intelligence" and "expert systems", I think that there are many good opportunities of using knowledge-based methods in statistics production. We shall return to this issue in chapter 4.

2.2 Survey operation vs survey planning, administration and evaluation vs strategical planning of the tasks and organization of a statistical agency

As was indicated above, we may look at the work of a statistical agency on three different levels. On the first, basic level, we have the actual statistical operations, making up the operational parts of a statistical survey. We all recognize the traditional, serial flow of tasks that have to be performed, when we conduct a survey: data collection, coding, editing, data transformation, aggregation, tabulation, graphical presentation, analysis, publication, distribution. Still existing, unused potentials for development of the computer-support in this area will be discussed in chapter 3.

On the second level, we control the different steps in the survey, and the survey as a whole, by means of design and planning, administration, and evaluation. The statistical design includes the establishment of a frame and a sampling strategy, if any, and the EDP design includes systems analysis, data modelling, and programming. The potentials for improved computer-assistance in these tasks will be treated in chapter 4.

Finally, on the third level, we look upon a statistical system as a whole as the object of control. The statistical system under consideration may be the statistical information system of a country, or a part of such a system that is managed by a particular statistical agency. Even though such a system will be based upon a number of statistical surveys, it will also contain other components, and it is true for statistical systems as for other systems that the whole system should be something more than just the sum of its parts. In chapter 6 we shall discuss in what ways a statistical information system can be something more than the sum of the surveys that it contains, and how a systems approach based on modern technology can help the supersystem to fulfill its purposes as effectively as possible.

3 Continued rationalization of survey operations through standardization, decentralization and integration

One of the principal messages of this paper is that we should actively look for possibilities to use modern technology in new areas and new aspects of statistics production, rather than just being busy moving "the same old applications" from one computer generation to the other. On the other hand, we must not neglect possibilities to do "the same old things" in a much better way by applying new technology and new methodology to the traditional tasks of the operation of a statistical survey. I will use three slogans to describe what needs to be done: standardization, decentralization, and integration.

3.1 Standardization of technology, software, and methodology

Computers are now so cheap, and people so expensive, that it is very rare that it is really worthwhile to aim at maximum technical efficiency in the design of a computerized information system. Naturally, in a big system, with large volumes of data to handle, and with a heavy traffic of man-machine interaction to take care of, it may sometimes be necessary, or at least clearly cost-efficient, to mobilize technical ingenuity in order to eliminate potential bottle-necks and speed up response times, or save storage and processing capacity. But there are two points to be made here. One is that there are not so many systems of this nature, not even in a statistical office with its large data bases. Most statistical surveys are small or modest in size, they are processed rather infrequently, and response time requirements are

often quite moderate in comparison with those of many commercial on-line systems of administrative character.

The other point I wanted to make in this context is that even in those few cases, where technical optimization is really optimal, from an executive point of view, it is usually not executively optimal to technically optimize the system as a whole, but only some limited part or aspect of it.

As a consequence of these observations, when it comes to the rationality of technical optimization, the burden of proof should always lie with those who claim that it is necessary. The default solution should always be a simple, straightforward, standardized solution.

What I have just said may seem to be so obvious that it need not be said. Unfortunately, this "obvious truth" does not always seem to be well understood or widely accepted in statistical offices. Even if we theoretically accept its validity, we seem to neglect its consequences when it comes to practice, not least when it comes to managerial practice. This seems to be a case where some active **delearning** of an old, puritan habit (not to waste any type of material resources under any circumstances) should be exercised among all those involved: programmers, systems analysts, and managers, to the benefit of the overall economy and usefulness of the systems designed.

In other words, a systems analyst or a programmer should never be allowed to deviate from the standards, set by the top management of a statistical office, without the hearing of a responsible manager, and a responsible manager should never accept a deviation from the standards without lengthy arguments, based on good documentation, the contents of which he or she fully appreciates. Violation of these rules should be regarded as a serious fault by internal and external auditors.

The proposed managerial rule assumes that there is a well worked-out standard, or policy, controlling important and tangible aspects of information systems design in the statistical agency. Ideally this policy should be well integrated with a theoretically sound methodology for systems design and implementation, so that the rules of the policy will be more or less automatically followed by anyone who uses that methodology, and its accompanying working tools (cf section 4.2). Objects for standardization are hardware, software, interfaces, and the methodology itself, including documentation rules.

3.1.1 Standardization of hardware and operating systems

For the time being the standardization of hardware and operating systems is not really a major problem, since the industry has (just by chance?) solved it for the users by establishing very strong **de facto** standards: IBM-compatible mainframes, IBM-compatible micro-computers, possibly UNIX. However, it should be noted and kept in mind that this good situation is not at all the result of some explicit action on the part of the users. Thus the scene may change rather quickly back to the more traditional lack of standards.

3.1.2 Software standardization

The most important phenomenon in the software area is that it seems to have become finally accepted that users should not develop their own application specific software. Instead they should use generalized software and, if necessary, customize these

products for their specific applications. As a result, the number of application programmers should decrease rather drastically in statistical offices, if it has not already started to do so. Some statistical offices have even started to question their own development of generalized software. Thus it seems that we are not far from questioning the whole professional category of programmers as an identifiable group of specialists within statistical offices. We shall return to this controversial issue.

It should be noted that when we talk about standardized software, or generalized software, in a statistical office, there are two important subcategories. One type of generalized software is the product which has been developed for particular statistical tasks, or functions that are typical for statistical offices. Such **generalized statistical software**, or **special purpose generalized software** has often been developed by the statistical offices themselves, or by institutes doing statistical research. There is another type of generalized software that we may call **multipurpose generalized software**. Such a product has typically been developed for general, commercial purposes, and not in particular for statistical applications, although they may also be useful in statistical environments. Data base management systems are a good example.

3.1.3 Standard methodology

Today most major computer-using companies and organizations in Sweden have adopted some standard methodology for the development and maintenance of information systems. The methodologies are called **systems development models**, and they say something about what tasks should be performed during systems development, in which order they should be performed, what concepts to use during analysis, how to visualize the concepts and the results of the analysis, which rules and standards are to be followed, which documents to prepare, etc. Typically there is one standard, or at least one variation of a standard per company. However, some kind of de facto standard seems to be emerging, even between the companies, including such features as separation between infological (contents-oriented) and datalogical (technique-oriented) phases, emphasis on conceptual modelling according to some "three-concepts-methodology" (objects, properties, relations [OPR], or entities, attributes, relations [EAR]) during the infological phase, and emphasis on the relational data model during the datalogical phase. As a complement to the state-oriented modelling of concepts and data, the systems development methodologies often contain flow-oriented modelling techniques and other methods for clarifying dynamical aspects of the system.

One important aspect of systems development methodologies is that they may prescribe **standardized interfaces**, for example standard formats for the storage and communication of data and metadata between different parts of a system, and standardized user interfaces in general, or standard syntaxes for user languages in particular.

The systems development methodology of Statistics Sweden is called the **SCB model** and has been in use for over a decade. It has also been the source of inspiration for the systems development models of several major (non-statistical) companies and organizations outside Statistics Sweden.

3.2 Decentralization of computers and computer-related resources

An obvious organizational consequence of the changes in cost relationships should be that there is no longer the same need for centralized control of computer-resources as there used to be. We are now able to buy such resources in small pieces and at a

very low cost per piece. Thus there is no need for centralization for the reason of sharing expensive, indivisible, and scarce resources. Instead we can integrate the decisions concerning computer resources with other important decisions in the statistical office and try to develop the same type of "balanced decentralization" of decision-making as in other areas, letting those responsible for a statistical survey take as full responsibility as possible for all types of resources needed for the design and operation of the survey, not treating computer resources in any special way.

In Sweden we have taken some important steps in this direction. The Government has initiated a process where those agencies which have enough competence and experience in EDP are allowed to take most computer-related decisions without having to ask any other agency or the ministry of finance. A condition for this freedom is, of course, that the agency is able to handle all decisions within its given budget.

Within Statistics Sweden we are also trying to treat EDP-related decisions, not separately, but integrated with other decisions. We have established an EDP policy, which will of course be updated from time to time, and within this policy, each department is authorized to take its own decisions, as long as they are within the budget of the department. Thus, as far as possible, all types of costs - for hardware, software, and personnel, for mainframe-related resources and for micros - are measured in "the same kind of money".

In the area of application systems design and programming we started the decentralization process already five years ago. The responsibility and the personnel resources for these activities were then removed from the central Systems Department to the different subject matter divisions. The remaining parts of the Systems Department were merged with some other development functions into a newly formed division for research and development and retained the responsibility for such functions as development and maintenance of generalized software, research and development in the area of statistical data processing, and EDP training.

Within each subject matter division the decentralization process has continued more or less rapidly. In some cases there are still some relatively big pools of systems analysts and programmers, and in some other cases the decentralization process has continued down to the level of individual subject matter units and surveys.

One good effect of this decentralization is that the manager of a particular statistical survey has now much more complete overview, knowledge, and control of his/her product and all types of resources that are needed, assuming of course that the manager has the capacity and willingness to make use of these opportunities. On the other hand there is naturally a risk that the statistical office as a whole will fall apart into a large number of small, uncoordinated survey based organizations. In order to prevent this, a number of specialized "councils" (among others one for EDP) have been created for giving specialized advice in policy matters etc to the top management and the Director General of the office.

In my opinion this decentralization process has by and large been successful and necessary and will continue in the future. Hopefully this will among other things lead to a better integration of different methodological aspects of statistics production, including the integration of statistical methodology and EDP. We shall return to this issue in the next section.

3.3 Integration of different types of tasks, skills, and competencies

Integration of different tasks, skills, and competencies is the other side of the "decentralization coin". The effects of this integration are becoming visible throughout the organization. Managers are losing their personal secretaries and administrative assistants, who find themselves replaced by personal computers, networks, and office information systems. Subject matter statisticians are taking care of application development and maintenance without the assistance of systems analysts and programmers, and they do some of their own publishing without having to rely on typographers. Interviewers take over data entry and data editing tasks.

Integration of these and similar types have been enabled and facilitated by the decentralization of computers and computer-related resources. The integration has a number of good effects, including job enlargement, shorter communication and decision paths, less administrative overhead, and a more clear division of responsibilities. On the negative side there is a risk of "happy amateurism" replacing competent and efficient professionalism, and of isolationism and self-conceit in the relatively small and independent organizational units. However, on the whole the positive effects seem to outweigh the negative ones, and there seem to be more staff members who feel they have gained from the development than who feel they have lost. Even some of those who have lost responsibilities and empires welcome or accept the development as being basically sound and find new roles in the organization relatively quickly.

Naturally, there will always be a need for good specialists in several fields of competence in a statistical office. The on-going technological development only eliminates a need for centralization and functionalization that was based on the indivisibility of large, expensive computers, and on the relative scarcity of systems analysts and programmers. However, the development has also raised the question, whether it would not be rational to aim at a higher degree of integration between the disciplines that are of relevance for statistics production. The division into specialties that exist at present in statistical offices by and large reflect the university organization. For a statistical office it would clearly be fruitful with a closer cooperation between statistical methodologists and computer specialists. Such an integration would be even more important, if it could also induce a change at the universities, making statistics production **per se** an established and respected area of academical research and qualification.

4 Computer-assisted design of statistical surveys and statistical information systems

No doubt, the next wave of computerization in statistical agencies will concern the environment of the statistical operations, rather than the statistical operations themselves. By and large we have already computerized the basic data handling and computation that is going on in the processing of a statistical survey. Now it is time to seriously consider a massive computerization of the tasks involved in the control of a statistical survey, tasks like planning and design of the survey from a statistical point of view, design and construction of the production system (which is in itself for the most part a computerized system), administration of the production activities, and evaluation of the performance of the statistical survey. Some of these tasks, and their computerization potentials, will be examined in this chapter.

4.1 Statistical design

In a statistical environment it is, of course, a well recognized fact that the statistical design is a most important activity in the planning and execution of a statistical survey. I am also sure that most statistical methodologists have already accepted the computer as an efficient calculator and tool in their design work. An interesting question is whether the computerization of the statistical design can take place in a more systematical and goal-driven way. The answer to this question can only be given by the statisticians themselves, in cooperation with computer specialists.

My personal belief is that statistical methodology, supported by an intelligent use of computers, could produce another round of rationalization of statistics production of maybe the same order of magnitude as the computers alone have already accomplished. It seems that non-statisticians sometimes think of statistical methodology as a "necessary evil", which has to be there in order to ensure reasonable quality of the results of a statistical survey, and to protect against conscious and unconscious misuse of statistics. Naturally this is an important function of statistical methodology, but statistical methodology in combination with modern technology could also be a very powerful, active force in the rationalization and cost-saving in statistics production.

Let me mention a couple of examples. In a major Swedish household survey, a team of competent statistical methodologists could, with the help of computerized analyses and simulations, propose a more efficient sampling and stratification strategy. The results from this exercise could be used for an improvement of the quality of the estimates produced in the survey. Alternatively the sample size could be reduced by some 50%, thus halving the data collection costs, a major budget item (about 2 MSEK) in this survey.

Another type of example is the significant gains that can often be made by a well designed combined usage of sample surveys and administrative registers. In Scandinavian countries this design could eliminate the need to take censuses, thus saving vast sums of money. An interesting characteristic of the approach is that it seems to be able to improve the quality of both the register and the survey, at the same time as it reduces costs.

Naturally a statistical methodologist could make this list of examples much longer and more precise. However, statistical methodologists do not seem to advertise and market the rationalization power of their methodological tools in this way very often. As a result, it seems to me that too many subject matter statisticians too often turn their interest to the latest novelties in computer technology, rather than seriously investigating some of the rationalization potentials offered by computer-supported statistical methodology.

4.2 Systems design

The growing usage of generalized software instead of tailor-made application programs has greatly improved the efficiency in systems construction and maintenance. However, the use of generalized software does not decrease the need for a good systems design. Nor does it in any significant way reduce the amount of work that has to be done by competent specialists during earlier design phases. It may seem surprising that exactly those specialists, who are themselves responsible for the design of so many computerized systems, do not to a greater extent use computerized systems to support

their own efforts. After all we have since long got used to concepts like Computer Assisted Education (CAE) and Computer Aided Design and Manufacturing (CAD/CAM).

However, right now something is happening in this area. A new acronym has been coined, CASE, standing for Computer Assisted Software Engineering, or **Computer Assisted Systems Engineering**. The CASE tool-boxes contain software instruments supporting the different working steps in systems development models. One problem is that each organization has its own systems development model, and there may not be any particular CASE tool-box on the market, which perfectly matches the needs implied by that model. The organization is then left with the alternatives of either having to change its systems development methodology, or to develop its own CASE tool-box.

As I mentioned in section 3.1, there is a process of de facto standardization going on in the area of systems development models. This in combination with the possibilities for an organization to acquire and customize a CASE shell (cf expert system shells), rather than having to accept all the details of a completely ready-made CASE tool, should help to solve the problem.

For a statistical office this development should be of great interest. While waiting for an adequate CASE shell to appear on the market, the statistical agency could itself undertake a number of relatively simple steps to improve the computer support in systems design.

One obvious, but important step is to develop an interactive tool for the creation and maintenance of systems documentation.

Another, related development step is to have a tool that automatically transforms and communicates metadata between different software products.

Furthermore, it could be questioned whether the designer/user should at all have to bother about more or less unimportant technical differences between different software packages, or even with the selection of a particular software product (rather than another one) in the first place. Ideally the designer/user should only have to specify the function (for example tabulation) that he or she wants to be performed, and then the systems development tool should automatically select (or propose) a software product and generate a complete application on the basis of metadata from the documentation system and some input from the designer/user, expressing his/her preferences on certain matters.

A documentation and systems development tool approaching the above-mentioned ideals has been developed at Statistics Sweden. It is called the CONDUCTOR and is running on the mainframe at present. It speeds up the work even of experienced systems analysts, and it makes it possible for people who are not computer professionals to get their own applications "in the air", provided that they have a relatively simple problem and/or an adequate understanding of the early, contents-oriented phases of the systems development model.

4.3 Knowledge-based methods

The term "knowledge-based system" is often used today as a more humble way of saying "expert system", which is in turn intended to be more down-to-earth than "artificial intelligence". Anyone who is not familiar with this jargon might rightfully question why we should suddenly need to start talking about using "knowledge-based systems"

and "knowledge-based methods" in statistics production. (Have we not always used methods based on knowledge? What other methods could there be?)

It **is** true, hopefully, that statistics production has always been based on knowledge, but typically this knowledge was not stored outside human brains, and if it was, it was usually stored on paper, separately from the computerized files, containing the data that were processed in accordance with the knowledge. And finally, if the knowledge **was** to some extent represented in a computer, it was usually stored implicitly in the programs.

In contrast, knowledge-based computer methodology assumes that

* the knowledge used in different parts of statistics production is (at least partially) computerized;

* the knowledge is organized as **facts** and **rules** in a so-called **knowledge-base**, which is handled in accordance with data base principles;

* the exploitation of the knowledge is actively computer-supported.

Thus, even though it is controversial, it must be admitted that one goal of applying knowledge-based methods to statistics production is to capture at least some small parts of the knowledge, which has up to now been regarded as inseparable from the statisticians who are in possession of the knowledge, and make it available to computers and to users of computers.

We should rightfully question to what extent this goal is a realistic one, but I think that we are not in a position to reject these ideas and proposals categorically. We must realize that we have only seen the beginning of a data explosion in society. Technically, anyone will soon be able to produce "statistics" from these data. But how can we prevent misinterpretation and misuse of this statistics production? The best thing would be, of course, if every amateur statistics producer would seek the advice of a competent and experienced statistician. But even if there were statisticians in such abundance that this would be a realistic possibility, I am not sure that most users would follow this path, and I am not sure that the competent and experienced statisticians would appreciate to spend 99% of their time giving routine advice on routine statistical problems.

I think that if we reason along these lines we can rather easily agree on a justified, desirable and realistic role for knowledge-based computer methodology in statistics production. In this perspective the knowledge-based systems is a natural step in the development, following the metadata systems and interactive user interfaces (cf the above-mentioned CONDUCTOR system) that we have already put into productive use, and which are appreciated by most of us.

Moreover I am rather convinced that the efforts to develop knowledge-based computerized systems for statistics production will generate some very good side-effects, even if the more ambitious goals should not be reached so easily. I think that we all agree that a statistical agency has its most important asset in their staff members, and in the competence and knowledge they possess. We have a big problem to maintain this knowledge capital when specialists retire, or when budgets are cut. Systematical documentation of the knowledge (called **knowledge acquisition** in the jargon of expert systems) could alleviate these problems and provide excellent instruments for in-house training.

5 The next generation of statistical software

When starting a discussion about the next generation of statistical software, the first relevant question is, whether there will be a next generation of statistical software, at least if we are thinking about software products developed by statistical offices themselves. Some statistical offices have already started to question the need for programmers and in-house software development.

I am sure that I am not the only one who would hate to see a statistical office without some competence in advanced software development. A disengagement in this area is probably an irreversible process, and it will have negative side-effects. For example, it may turn out to be difficult to evaluate, select, install, and adjust commercial software to the specific needs of a statistical office, if the office does not have a critical mass of competence in software development.

On the other hand it must be admitted that it will be increasingly difficult for statistical offices to justify glamourous software development projects of the costly type that we used to launch in the past, and which we sometimes (but not always) managed to complete and implement successfully.

There are some actions that can be taken in order to come to grips with this difficult situation. One is to establish a policy and basic architecture for software development, ensuring important features like cohabitation possibilities between commercial packages and in-house developed software components, modularity and incrementality in software development, and portability between different types of computers.

Another possible action, which should be combined with the first one, is to rely more on international cooperation between statistical agencies. I will use some experiences from the UN/ECE Statistical Computing Project (SCP) as a basis for the discussion of these matters.

5.1 International cooperation: The UN/ECE Statistical Computing Project (SCP)

SCP is an acronym that symbolizes a cooperation effort in the area of statistical computing between the countries of the UN Economic Commission for Europe (ECE), including the European countries, Canada, and the United States of America. SCP has been going on in various forms and with various themes of cooperation throughout the 1980s. It started as a project (SCP-1), supported to some extent by the UNDP, then it became a programme (SCP) under the Conference of European Statisticians, and since about a year and a half it is once again a project (SCP-2), supported by the UNDP. Whereas SCP-1 was basically mainframe-oriented, SCP-2 should pay special attention to the growing use of micro-computers in statistical offices.

The substantial work in SCP-2 has been organized into six Joint Groups. They are:

* the Joint Group on Software Evaluation
  (lead country: Hungary)

* the Joint Group on Communication
  (lead country: France)

* the Joint Group on Implementation Strategy
  (lead country: Poland)

* the Joint Group on Statistical DataBase Management
  (lead country: Sweden)

* the Joint Group on Data Editing
  (lead country: Yugoslavia)

* the Joint Group on Table Generation
  (lead country: German Democratic Republic)

The work of the Joint Groups is monitored by a Task Force, consisting of the lead countries of the Joint Groups, and the Task Force reports to a Steering Committee consisting of all countries participating in SCP-2. The three last-mentioned Joint Groups in the list above are actively engaged in software development in their respective fields of interest.

5.2 Some desirable properties of the software

In an attempt to amplify the total effect of the software development going on in the different Joint Groups of the Statistical Computing Project, the author of this paper was asked by the Task Force of SCP to come up with some concrete coordination proposals. Ideally this effort should result in

o a description of a unified design approach and software architecture to be shared by all the Joint Groups in the continued software development

o a proposal for a unified way of handling metadata in SCP software products

o a proposal for ensuring easy import/export of data between different software products (SCP software, commercial packages, homemade products etc) and between different computers

o a proposal for ensuring portability for SCP software between different types of computers (micros/minis/mainframes)

o a suggestion of steps to be taken to facilitate the "marketing" of SCP products as members of one and the same software family

o a tentative, synchronized plan of activities to be carried out by the individual Joint Groups in order to fulfill the common goals of the SCP software development

Some of the proposals from this mission will be presented in the next section.

5.3 A proposed architecture for the software

The report put forward as a result of the coordinative effort mentioned above includes the following proposals:

[1] All SCP software development should be based on a data base oriented model of statistics production. The elementary operations of this model should be carefully defined, and the definitions should be

based on a logically stringent, functional analysis of the typical major functions in statistics production. All intermediary operations in a production chain should use and produce data base objects of one and the same type: flat files, or relations in the sense of the relational data model. Thus in a mathematical sense, the operations would constitute an algebra over this type of data structures. This conceptual basis for the SCP software development will ensure modularity, simplicity, combinability, and incremental developability.

[2] In addition to portable and well integrated basic software components for editing, data base management, and tabulation, the SCP software package should contain a user-friendly Systems Development Environment (SDE), consisting of an Interactive User Interface (IUI) and, if possible, some tools based on the principles of Computer-Aided Systems Engineering (CASE) and knowledge-based methodology (expert systems). Among other things, the IUI should help the user to overcome any differences in the user languages that may exist between different SCP software components, and that will certainly exist between the SCP software package and other software products that the user may want to combine with the SCP software. Thus the IUI should help to standardize the user interface despite inevitable differences between software products. On the other hand, the IUI could also be used to individualize (customize) the user interface to fit the particular needs of a particular user group, or a particular statistical office.

[3] The uniformity of the SCP software products on the conceptual level should have a natural counterpart on the technical level. An algebra of operators working on standardized data structures (flat files and relations) has already been successfully implemented, and these principles should be generalized and applied to the other SCP software projects as well. Here a processor concept is proposed as the software technical counterpart to the algebra operators. Each SCP software component should be designed in terms of processors, and processors with identical or similar tasks should be standardized, and implemented only once. Thus identically the same processor could be used in several parts of the same software component, and in several software components.

[4] The processor language, controlling the operations of the processors, could sometimes also be the user language. For example, this is the situation in the case of the Base Operator System. For more complex functions like editing and tabulation, there is good reason to have a special user language that is mapped (translated) into the processor language.

[5] In order to standardize and facilitate the accessing and communication of data within processors, between processors, between SCP software products, and between SCP software products and the outside world, a common Relational Access Method (RAM) is proposed to be implemented. This access method should be used by all processors in all SCP software products for the reading and writing of data. RAM should consist of a set of independent macros, which can

be included in the processor modules. In order to connect an external software package to RAM, one will have to develop the appropriate read/write macros etc for the particular package, but this will be a relatively minor task.

[6] Similarly, in order to standardize and facilitate the handling of metadata, including - as far as possible - an automatic transfer and transformation of metadata between processors, between SCP software products, and between SCP software and external packages, a common Metadata Management System (MMS) is proposed to be implemented. MMS should be used by all processors in all SCP software, and it should consist of a set of macros for the reading, writing, updating, deleting etc of metadata. Using such a set of macros, it would be quite easy to support different types and formats of metadata without any changes in the processors. However, it should be noted that this is a proposal for a standardization of the handling of metadata, not for a standardization of all kinds of metadata, which does not seem to be a realistical objective at present.

[7] In order to ensure maximum portability of SCP software between different categories of computers and operating systems, the programming language C is proposed to be used in all software development. The portability should (with priority 1) be certified for IBM PC compatible micros under PC/DOS and MS/DOS, for IBM 370 compatible mainframes under OS/MVS and VM/CMS, and for the operating system UNIX.

[8] The design, implementation, and documentation of all SCP software products should cover the following items, in the following order:

* systems analysis and formal description of the particular statistics production function under consideration

* development of a reference manual for the end-user language

* module specification in pseudo-code

* development of a systems manual for the software system under consideration, containing a description of

  - the logical program structure

  - the processor language statements that are used and produced by the software

  - the possibilities to write tailor-made exits

  - other features for an advanced use of the software

* coding in C

* preparation of

  - a user's guide, based on pedagogical examples

- an installation guide

- an installation tape and installation diskettes, containing the software and its documentation

- additional documentation if necessary

[9] In applicable parts, the design, implementation, and documentation of common SCP software components, like the processor language, the Relational Access Method (RAM) and the Metadata Management System (MMS), should cover the same items as listed above.

5.3.1 Some comments to the proposals

The proposals presented above are on a relatively high level of ambition, and of course it remains to be seen to which extent they will be accepted and realized by the SCP Joint Groups. In particular it may be noted that the proposals assume that the software development will be carried out in the programming language C. An alternative could be to base most of the software development on program generation techniques and/or some commercially available, portable, general-purpose software product like the database management system ORACLE with the standardized interface SQL. Most of the other proposals in the list above would be relevant anyhow, but naturally they have to be reinterpreted to some extent.

6 The future architecture of statistical information systems

6.1 Is there a need for a statistical office any longer?

It was noted in chapter 3 of this paper that the technological development has alleviated one restriction on statistics production that used to exist: the necessity to share scarce and expensive computers and computer-related resources. This has started a decentralization process. Are there any natural limits to this development, or will the decentralization stop only when the statistical office has been dissolved into a number of separate statistical surveys? In other words: will there be a role to play for a statistical office as a strong, independent organizational entity in the future, or will the statistical surveys be taken care of by other governmental agencies?

Personally I do believe that statistical offices have an important role to play in society, quite regardless of the decentralization possibilities that the technological development is now offering, but I think that we need to ask those critical questions indicated above. Others will do it.

Statistics production in Sweden was centralized into its present form in the early 1960s. The need to rationalize efficiently by means of centralized computer technology was then a major reason for centralization. But there were also others with at least the same dignity. One was the belief that only a strong, central statistical office could afford to maintain a powerful methodological development of high quality and enough quantity to form a "critical mass". Another reason for centralization was the needs for coordination and integration of individual surveys into statistical information systems, based on unified conceptual models like the system of national accounts and socio-demographical and socio-economical models.

Until recently the technological arguments for a centralized statistics production have been so widely accepted that we have not had to use the other, more sophisticated arguments. Maybe, as a consequence, we have not been so active in the areas of methodology and integration as we should have been.

6.2 User needs

The needs for coordination and integration are deeply and directly founded in some strongly felt user needs. For example, the statistics customers of Statistics Sweden are rightfully irritated when they have difficulties to locate and interpret the statistical data that they are looking for, and they are equally rightfully irritated when they have to go to several places in the organization in order to get all the data they need, instead of getting everything in one place, including some advice about how to combine data from different sources.

Furthermore there is a growing number of rather advanced users of statistics, with more or less sophisticated models and hypotheses that they would like to try on official statistical data, and sometimes combine with their own data. Due to the technological development these users will always have access to powerful computer equipment of their own, and they have a good understanding of the possibilities offered by modern technology. If these users are not well served by the statistical office, they will exercise all the rights that they may have to obtain statistical data in rather "raw" form from the statistical office, and use them together with their own data, software, and models in **data laboratories** that they build and run independently of the statistical office. If a statistical office wants to be successful in this competition it must be active, imaginative, and flexible, and it must use its relative advantages in methodological competence, and coordination and integration possibilities.

6.3 Needs for rethinking?

The statistical survey is the basic building block in many statistical organizations. I have pointed out that the on-going decentralization will further strengthen the power and control in the hands of individual statistical surveys. From a managerial point of view, this development is good in many respects. It clarifies responsibilities within the organization, and the person who is in charge of a statistical survey will not have so many others to blame, if something goes wrong.

On the other hand there are those user needs discussed in the previous section, which call for other organizational solutions. In order to make it easy to locate and interpret data, all statistical data of any importance must be well documented, and they must be documented in the same way from survey to survey. Thus all statistical data must be accompanied by appropriate **metadata**. The metadata must be computerized, and like the data themselves they must be organized in accordance with **uniform database principles**. Many statistical offices have since long been trying to implement these ideas in different ways, but the results are not always encouraging. In today's competitive situation it will not be enough for some surveys and some departments to be engaged in this work. Instead a systematical implementation throughout the organization may be a question of vital importance for the statistical office as an independent entity.

Databases and metadata will not be enough to serve the needs of the users. With all data and metadata easily available, they are certain to **combine data from different** sources, that is, data emanating from different surveys and other sources, like

administrative registers. And they will make such combinations whether we approve of it or not. Traditionally, statistical offices have been able to hide behind their publications. A statistical survey is responsible for the contents of the reports and publications that it publishes, but it takes no responsibility for how the user may possibly combine the data in the publication with data from other publications. This strategy will not be maintainable in the era of new technology.

This is a difficult problem, and there is no simple solution to it. However, it seems clear that statistical offices must activate themselves in the area of **standardized concepts and classifications**, an area whose importance is not always fully appreciated by those in charge of individual surveys, and sometimes not even by managers on higher levels. On the other hand this seems to be an undisputed area in the sense that most people outside a statistical office accepts the office's responsibility and welcomes its competence in this area.

Unified concepts and classifications is an excellent basis for combining data and putting them into models. However, there will always be cases where complete standardization is not possible, and it is important that a statistical office plays a constructive role also in such situations, even though it may not itself be responsible for the difficulties. For example, due to the different purposes of an administrative register and a statistical survey, it is inevitable that all definitions cannot be harmonized between them. However, experts in statistical offices should take as their responsibility to find ways around the problems, exploiting in a positive way the power of statistical method.

By means of these examples I have shown that there is a need for managerial action and control that is **global** in relation to the individual surveys. One may go one step further and say that the new problems and possibilities call for a **new survey concept**. Traditionally, a survey has been modelled as a serial flow of production steps, starting with survey design and data collection, and ending with tables, reports, and analyses. One effect of modern technology is that the ties between input and output will be weakened, both physically, logically, and in time. The statistical end-products and typical usages of statistics will be based on combinations of input data from many different sources, and the data collected by one statistical survey will be used for many different purposes, by different users, and at different points of time.

Thus, if we look upon a statistical survey as a basic building block of statistical organizations and statistical information systems, it may be more adequate to think in terms of three different types of surveys:

* **input-oriented surveys**, collecting and editing the data, performing some routine tabulations and analyses, and preparing the data for future use by putting them, with their accompanying metadata, into common databases;

* **common databases**, taking care of data from different input-oriented surveys, and forming the basis for output-oriented surveys;

* **output-oriented surveys**, making use of existing data in common databases and other sources, inside and outside the statistical office.

THE USE OF NEW TECHNOLOGIES AT STATISTICS CANADA

J. E. DOUCET,

Statistics Canada

THE USE OF NEW TECHNOLOGIES AT STATISTICS CANADA

October 24, 1988

J. E. Doucet
Informatics Branch
Statistics Canada
Ottawa K1A 0T6

Abstract

This paper presents a survey of the new technologies which are (or might soon be) available at Statistics Canada. Interesting current applications of new technologies within the organization are identified, and the paper speculates on the future direction of these seminal applications. Finally, the set of new technologies which have not yet found application at Statistics Canada is examined, and the author speculates on their suitability for eventual integration into the set of technologies to be exploited within the organization.

Introduction

During the past twenty years, we who work in central statistical offices have witnessed successive waves of new computer-related technology. Each wave has brought significant increases in power and functionality - at a steadily decreasing cost per unit of computing. These quantum improvements in technology have driven our efforts to automate; and the degree of automation has been substantial. At Statistics Canada, we have been able to shrink (gracefully, I think) from a staff of approximately 6,000 employees during the late sixties, to a staff of about 4,400 today. And we are visibly more productive as an organization, than ever before.

However, automation has sometimes been a mixed blessing. We have had to tolerate disruptive conversions from one plateau of technology to another; some of us have had to wrestle with the complexity and number of man-machine interfaces which we have had to learn - in order to perform our jobs; and we have become well practised at solving the same problems over and over again as each new wave of technology has been introduced and applied.

Today, managers are begining to take a much more strategic view of new technology. Their lessons learned, they are being more selective in their choices of specific technologies, in at least two related ways:

- they are demanding that specific new technologies be easily integrated into their particular local fabric of computer-related services; and

- they are demanding that any new acquisitions comply with the several major international standards which have gained prominence over the past few years.

Certainly, this is now the case at Statistics Canada.

A Local Framework for Assessing New Technology

During a recent planning exercise at Statistics Canada, the concept of an "integrated computing environment" (ICE) was advanced. This ICE is defined as one which offers a stable, consistent and manageable set of computing services, in a relatively seamless fashion, across a wide range of hardware systems. An important feature of the ICE is that any computing service may be accessed from any workstation in the system.

Such an environment is desireable because it provides users of computer-related services with one consistent set of man-machine interfaces, rather than a different set of man-machine interfaces for each of the various computer platforms which they must use. There are other benefits:

- system support activities may be focused on a smaller set of functionally integrated software products, to the benefit of all users;
- applications are able to migrate from one computer platform to another in response to operational needs, differentials in rates, etc.;
- significant hardware savings can be realized from the ability to operate more and more applications from virtually any hardware platform - allowing competitive bidding for hardware acquisitions; and,
- volume discounts on software licences can be significant, as more copies of a smaller set of products are purchased.

The ICE presents an interesting paradox. In the past, stability and integration have generally been considered to be at odds with innovation. Today, it can be argued that the more stable and integrated the computing environment, the better it should be able to react to technological innovation; and the better positioned it will be to integrate the results of that innovation into its overall service offering. Why? The underlying reason is that the ICE is necessarily based on industry-wide and international standards: SQL for data management, C-language for third-generation compilers, the OSI-model for interprocess communication, UNIX (and a small number of others) for operating systems, etc. Increasingly, technological innovation is taking place within this broad suite of standards (official and de facto) in order to be commercially viable. Thus, by remaining a very mainstream affair, the ICE is well positioned to benefit from the fruits of technological innovation.

The Move to an ICE at Statistics Canada

At Statistics Canada, the major elements of an ICE have been put into place over the past two years.

- In the key area of data management, a commitment has been made to the set of relational technologies - especially SQL, which will soon become a Canadian Government standard for data management software acquisitions. Statistics Canada chose the ORACLE product over two years ago because it can be offered across all of the organization's computing platforms.
- The operating systems are being sorted out - slowly but surely. The growth in the number of proprietary minicomputer operating systems has been halted. And, the systems with a future at Statistics Canada appear to be: MVS, DOS and UNIX. The OS/2 operating system will likely be limited to LAN file-server use - at least, in the short term. Where other operating systems exist, the more proprietary features of those systems are being avoided.
- A standard communications infrastructure is slowly being assembled. It features - Ethernet, TCP/IP, SNA; with an eye to eventual OSI migration. System interconnection is increasingly via Ethernet Gateways, rather than bi-lateral protocols or emulation packages.

- The default third generation language is C-language, because it can be offered consistently across all processors. It is displacing PL/I for new applications - and it is being used extensively for systems development using microcomputers.

- Proprietary (and locally developed) statistical software is increasingly of the functionally integrated type, which can be offered across computing platforms; for example, the SAS software is available on the three classes of processors, and the new suite of locally developed survey processing software is being made fully portable.

The ICE is evolving; but it must currently coexist (and sometimes overlap) with elements of the old environment, which cannot be easily integrated in this fashion. This will be the case for another 15 years. However, the older unintegrated services will progressively be displaced, in favour of the ICE approach[1].

Using the ICE

It is against this model of an integrated computing environment which new technologies must be evaulated. If a new technology does not fit the model, one of two paths must be followed:

- the technology must be rejected; or

- the model must be examined, as a whole, to see if it might be changed to accommodate the new technology.

Expediency has no place in the ICE model; a new technology is never acquired on the basis of its ability to best meet the local needs of a single application. It must fit - or it must be made to fit.

With the ICE as a reference model, it is possible to take the measure of the newer technologies which are at issue today.

SQL and the Relational Technologies

SQL is probably the single most important of the new technologies - because it has the potential to integrate applications above the level of the specific database engine. SQL is now viewed by proprietary software vendors as the defined standard interface between their products and all other relevant products. One important example will illustrate: the SAS Institute (a company which is almost completely market-driven) recently announced full SQL support for the SAS family of software, making it possible to underly SAS applications with relational databases!

The relational approach itself has long been of interest to statistical offices - because of the nature of our applications. At Statistics Canada, the General Survey Function Development (GSFD) is an interesting case in point. GSFD is the name given to a wide-ranging set of technical strategies for re-implementing the suite of generalized survey processing software at Statistics Canada. It features a strongly relational approach and threatens to redefine survey processing.

The GSFD Example of a Relational Approach

The following diagram may be used to illustrate the first level of functional decomposition for the GSFD design.

There are five major functions - all linked through the services of a full featured relational database management system (RDBMS). The GSFD would not be attempted if this generation of data management software were not available. Its power and simplicity are key to the success of the system.

---

[1] In some cases, older products can be adapted for the ICE. For example, it may be feasible to adjust the RAPID data management software to the extent of providing it with a SQL-interface. This would make some SQL-based third-party retrieval software available to the Census application.

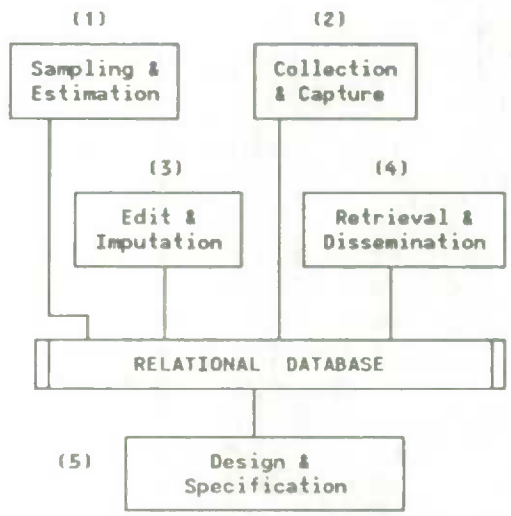


Figure 1: A conceptual view of the inter-relationship of the major components of the GSFD.

Function-5, in the figure, may require some explanation. It represents a system-wide design specification tool. Initially, each function is being individually specified. For example, the Edit and Imputation function currently accepts (interactively) a specification format which is similar to that used for linear equations. Ultimately, individual specification functions will disappear, to be replaced with a comprehensive design specification function which will operate at the level of entire survey specification. This will be mentioned later, as an example of the potential application of Expert System technology.

The relational approach permits the suite of software to be database-driven. The designer will be able to reposition specific functions from one component to another. For example, a set of edits may be moved forward or backward in the processing chain - from preliminary editing in the Regional Offices to post-imputation re-editing on the mainframe, perhaps. As a further example, deterministic imputation may be specified for execution at the time of preliminary editing in the Regional Offices, rather than later in the processing cycle - as is currently the case. In summary, specific functions may be deployed to suit individual application needs. Thus, the old time-sequential model of survey processing becomes only one of many models which the GSFD will support. The model may be constrained in different ways; no individual set of constraints is intrinsic to the model.

More generally, there is the potential to share valuable utilities across GSFD functions. For example, the Automated Coding utility which is currently available on the mainframe (ACTR - Automated Coding using Text Recognition) is being retooled to integrate into the GSFD environment. It will be driven by the same relational technology as the rest of the system. Thus it will be possible to provide access to the significant benefits of automated coding, to any GSFD function. Another utility which is being retooled is the one which provides full geocoding for any Postal Code. Inevitably, others utilities will follow.

At a more technical level, future versions of SQL will permit site-specific operators to be added to the language. For example, it will be possible to design an SQL operator for calculating "length-of-stay" in an institution. Over time, the inventory of such standard algorithms could grow dramatically - and systems development would become primarily a problem of assembling an appropriate solution, and adapting it to a specific operating environment.

The Relational Connection to New Hardware Technologies

The importance of the relational model has been recognized at Statistics Canada (and elsewhere) for a long time - on paper. However, it is only during the last few years that the computing hardware has undergone the technical revolution which was necessary to make relational database software feasible. Relational software needs to be able to manipulate very large tables, in memory and at high speed; consequently, it needs very fast CPUs and lots of cheap memory. We are now witnessing the first wave of computer systems which are adequately configured for relational database software. Other

waves will come - and they will be highly complementary to the needs of relational database software: parallel processors, vector processors, better virtual operating systems, 32-bit operating systems for large micros, bigger and cheaper memories, etc.

The Relational Connection to Distributed Database Technology

With an RDBMS, the repository of data is logically unique; physically, it is likely to be spread across several systems. The current state of development of relational database management software does allow for individual tables of a relational database to reside on different CPUs - transparently to the user. The DBMS software knows the network and it assembles the data for the user. Alternatively, different portions of the logical database may reside on different CPUs - but not transparently. Initially, Statistics Canada has opted for the latter because its network is not mature, and it has the special problems posed by the specific security and confidentiality concerns within its network policy. Nevertheless, true distributed databases remain a technical option (currently unexploited) which is consistent with the organization's ICE as currently defined.

It should be noted that meta data also reside in the relational database. For example, all of the design and specification intelligence from Function-5 in the GSFD example can be expressed in tabular format and kept directly accessible from within the same logical database. In the specific example of Function-2 (Collection and Capture), this type of information may be required for the production of customized data collection documents - a distributed function.

Expert Systems Technology

In many organizations, there exists a deep psychological need to do something useful with existing Expert Systems (recently downgraded from Artificial Intelligence) tools. Statistics Canada's commitment to date has been modest - and for good reason, I believe. While some statistical offices have undertaken some interesting small demonstration systems, the potential benefit of this technology remains effectively unexploited.

This year, Statistics Canada will take its first tentative steps into this area of new technology. The first application will likely involve the building of an intelligent front-end for the Generalized Iterative Record Linkage System. It is an application which boasts a genuine "expert"; and the expert service is currently over-demanded. A more ambitious undertaking (which may proceed in parallel) involves the problem of survey design/specification which was outlined earlier in the GSFD example.

It is important to integrate Expert Systems into the mainstream of computing. For that reason, we are undertaking development on UNIX-driven Sun workstations using industry standard PROLOG and supported by the same ORACLE relational Database technology found elsewhere in the ICE. This is crucial, if the results of Expert Systems research are to be eventually ported from enriched development environments, to less sophisticated environments for production use.

Expert Systems applications are typically difficult to delineate; they are often buried in the bowels of a much larger - and more conventional - application. Thus, there is a clear need to restrict the use of Expert Systems tools to those which can be fitted into the overall informatics service offering.

Workstation Technology

There is nothing unique about the use of workstation technology at Statistics Canada. The population of micro-based workstations now numbers approximately 1,400. As elsewhere, the standard workstation is slowly being redefined upwards to the 80286-based level and beyond.

Initially, these workstations tended to operate in a stand-alone mode, or to be clustered in small Local Area Networks (LANs) in order to share the cost of large disk storage systems and laser printers. More recently, the increasing power of the machines (and our familiarity with their capabilities)

has led us to attempt more serious forms of data processing with them. Consequently, we are now witnessing changes to the simple initial environment:

- the introduction of powerful file servers designed for LAN application;
- the introduction of LAN versions of relational database products - replacing single-user, non-SQL systems;
- the upgrading of communications protocols from XNS-based to TCP/IP-based in order to facilitate communication with other computing environments;
- the integration of multi-user 80386-based systems (UNIX, typically) into conventional Ethernet LANs.

All of this is now taking place, at a pace dictated (not by the technology, but) by our ability to pay for the desired changes.

Interesting uses are being made of microcomputers at Statistics Canada. For example, the entire Edit and Imputation component of the GSFD was developed on micro-based workstations (first under DOS, and later under UNIX) using all of the same software which would have been used if development had taken place o n the mainframe. Porting to other environments has been easier than anticipated - but not trivial. An additional benefit of this style of development, has been the ability of developers to offer to other members of the development team (Subject Matter Specialist and Methodologist) the opportunity to take the system back to their offices - to their own micro workstations - and kick the tires at their pleasure. This is a very useful mechanism for determining the true requirements of an application at the earliest possible time.

Secondary-Storage Technology

The most interesting advances in this area of technology are probably represented by the newer optical disk products.

The more familiar of these is the Compact Disk - Read Only Memory (CD-ROM) technology, which has been exploited primarily for the distribution of CANSIM products. CD-ROM applications are generally identified with a requirement to distribute many copies of a product, when the time constraints are not critical. Looking at this statement another way; CD-ROM products exhibit a high initial overhead cost, and the mastering process implies time delays.

The lesser known variant of this technology is the Write Once Read Many (WORM) approach, which is currently the subject of a research project undertaken by the International Trade Division and Research and General Systems of ISD. WORM is better suited for low-distribution, time-constrained applications. Again, looking at this another way; WORM products have a low start-up cost, but a high per-copy cost. However, no mastering process is required, so the product can be used in time-sensitive applications.

More recently, laboratory versions of the WORM technology have been developed, which are eraseable[2]. This development would bring optical storage into direct competition with conventional disk technology; and this would be the impetus required to adapt standard access mechanisms to support the optical technologies. To date, the capacity of conventional disk technology has been increasing at such a rate that only the arrival of the eraseable WORM would threaten its supremacy.

Conventional WORM devices are currently non-standard, and operate outside the ICE scheme. Still, they can certainly be exploited in limited ways:

---

[2] Obviously, the WORM acronym will have to change when these products become commercially viable.

- as local storage medium for various dictionaries: the SIC manual, International Disease Codes, etc.;

- as a low volume distribution medium for items normally making use of micrographics (fiche);

- as a throw-away backup mechanism for micro-computers; etc

At such time that their use becomes more standard - particularly with regard to their data management technology - their 'goodness of fit' into the ICE can be re-evaluated.

With conventional storage technologies, Statistics Canada has travelled the same road as most other large organizations: higher densities on all media; better packaging (cartridge tapes, for example); and continually increasing total storage capacity. Robotics and Solid State Disk remain areas to be studied further for possible investment.

Electronic Imaging

At Statistics Canada, Electronic Imaging exists in two separate worlds. Within the ICE, Laser Printing systems are available - in the LAN environment, in the mainframe environment (IBM 3800), and in the class of "departmental" systems (IBM 3820). These products are continually advancing, and the ability of the software to integrate text and graphics is improving all the time. Recent examples can be seen in the evolution of the mainframe's GML (general markup language) processor. We have recently introduced Formula Processing[3] plus new tags which support in-line graphics for downloading to some printers.

The other world of Image Processing is quite distinct from the ICE, and is generally based in Xerox technology. It includes the 'serious' image processing more closely related to photocomposition. At the moment, there is no elegant interface between the two worlds. One of the remaining challenges to the concept of an ICE is the building of convenience 'bridges' between these two dissimilar worlds of technology.

Desktop Publishing is the remaining fragment of this area of technology. At the moment, it may be compared to the situation for Optical Disk: the standards are not in place, and it should be restricted to peripheral uses of the technology. It does not integrate easily.

Specific Hardware Advances

As was discussed earlier, many of the hardware advances appear to be closely related to the exploiting of the new relational technologies. In many respects, the advent of relational processing software is attributable to the progress made in hardware related development:

- availability of increasingly larger computer memories,

- faster CPU's,

- virtual memory Operating Systems with the ability to address very large spaces,

- very fast disk (including solid state disk),

- parallel processors, vector processors, and array processors.

The extent to which non-relational software can be in a position to benefit from most of these advances is moot. The ICE which has been defined as a model for Statistics Canada is well positioned to benefit.

In 1986, Statistics Canada did conduct an elaborate evaluation of vector processing. In brief, a representative statistical processing workload was assembled, and arrangements made to run it on three different architectures of vector processors:

---

[3] This Chi Square Distribution is an example of formula processing using GML -

$$F(\chi^2) = \int_0^{t^2} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n-2}{2}} e^{-\frac{x}{2}} dx$$

- IBM's 3090 - a processor which integrates vector processing in the form of additional hardware instructions - the NAS computers use a similar approach;

- CRAY's XMP - a standalone vector processor; and

- AMDAHL's VP-1200 - an attached vector processor working under the control of AMDAHL's scalar 'host'.

In brief, the results showed that the savings could be very dramatic, but that the tools did not integrate well with our concept of the ICE. For example, most of the tools were FORTRAN based, and there was no support for SQL. It was therefore decided to wait until vector processing moves closer to the mainstream of informatics processing. This is consistent with our approach to many other new technologies - including parallel processing, and array processing. Our ability to integrate them is paramount. For those which we cannot integrate, it is too soon! It is preferable to wait for them to come to us when if they become commercially viable.

Summary

The foregoing has highlighted the major new technologies which are currently of interest to Statistics Canada. It has also attempted to make a case for the ICE framework against which all new technology must be locally evaluated. Many smaller areas of technology have not been specifically addressed:

- CASE technologies which are just now being integrated into the product lines of the vendors of relational database software;

- Automatic Translation software, which can clearly be permitted to exist on the periphery of the ICE, without penalty;

- analytical workstations which use advanced interactive graphics; etc.

However, the mechanism by which these technologies would be evaluated for eventual integration has (I hope) been clearly established. Reference models like the ICE can play an important role in saving us from a future situation where:

- we have one of everything; and

- nothing integrates!

THE USE OF TECHNOLOGY AT STATISTICS SWEDEN

L. LYBERG AND P. DEAN,

Statistics Sweden

1. Introduction

Statistics Sweden is the national statistical office in Sweden. The agency produces most of Sweden's official statistics. Statistics Sweden has been commissioned to conduct, on behalf of central and local authorities, surveys and to provide an advisory service on statistics. The agency may also undertake commissions for private persons, corporations, and organizations.

There are four subject-matter departments: labor market statistics, area statistics, enterprise statistics, and individual statistics. In these departments over 200 surveys are conducted each year. Departments for statistical services, data collection and processing, and research and development serve the subject-matter departments. The current number of employees is 1669 which is a decline by a few hundred compared to a decade ago.

Severe budget cuts in the early 1980s brought about a relatively speedy development of what might be termed high technology. It also became more and more difficult to hire staff willing to do some of the traditionally manual survey work. For instance, Statistics Sweden used to hire temporary personnel for census data processing. The nature of the work and Swedish labor regulations eventually made it impossible to hire temporary labor and the agency had to rely on its regular staff for censuses too. This would not have been possible without the use of technology. Examples of new procedures first introduced in the censuses are optical character recognition (OCR) and automated coding. OCR replaced keypunching and automated coding took care of more than two thirds of the occupational coding.

All new technology has not always been the result of budgetary cuts or other rationalizations. Opportunities to do more things and new things have led to the use of new technology. This is the case with our attempts at developing a hand-held computer for data collection and the technology available for various forms of land-use statistics. During recent years Statistics Sweden has purchased several hundred PCs and the EDP environment has changed fast. One can definitely declare that survey work is very different from what it was, say, five or six years ago.

In this paper we provide some examples of, rather than an overview of, the use of technology at Statistics Sweden. First, the EDP environment is described. Then we describe our attempts to computerize data collection and coding. We end with notes on technological aspects regarding editing, privacy and confidentiality, land-use statistics, and data dissemination.

2. EDP at Statistics Sweden

To most people's minds, the very term high technology is synonymous with computers. Computers stand as the symbol of high technology in an age dominated, even defined by such technology.

so the capabilities of any statistical office is intimately linked to its computer power and how well that office can utilize the computer power at its disposal.

Statistics Sweden's computing environment consists of two mainframes (an IBM 4381-3 and a NAS AS/8053, a few minicomputers (IBM S/36 and UNIX), and about 300-400 personal computers. About 50 of the PCs are part of a number of different experiments with PC networks.

Statistics Sweden has two sites, and the mainframe computers are located in Örebro (a city about 200 kilometers west of Stockholm) and communicate with the terminals in Stockholm via leased telephone lines. The IBM 4381-3 has a processing capacity of 2x2.5 million instructions per second (MIPS) and a primary memory of 32 megabytes. The NAS AS/8053 has a capacity of 7.0 MIPS and a primary memory of 32 megabytes. Our storage systems include an IBM MSS 3851 that can store 169 gigabytes and a disk system that can store a total of about 50 gigabytes. We are planning to replace our mass storage system (MSS) and are considering a system with a robot-run tape archive.

2.1 Personal computers

At the time of writing, Statistics Sweden has 300-400 personal computers and this number is expected to double in the near future. At present, PCs are used mostly for office automation, for example, word processing of routine materials like correspondence and reports, and even word processing on a more sophisticated level, like desktop publishing with complicated layout and graphics. Other areas that fall under office automation are the use of spreadsheet programs like Lotus 1-2-3 for budgeting and for calculations that one would previously have used a hand calculator for. Most of the data analysis that is done at Statistics Sweden uses PC SAS. Data capture was once done on special equipment, CYBERDATA, but this too is now done on PCs.

Given current technology, smaller surveys can be processed, totally or in part, on a PC. Earlier studies have shown that over 50% of Statistisc Sweden's statistical products process less than 10 thousand observations and can be produced entirely on PCs. This means that in the future, the most important work for the mainframes will be managing databases and processing large surveys like the Census of Population and Housing. Nevertheless, it is planned that certain parts of the large surveys can be done on PCs.

There are three types of personal computers at Statistics Sweden: Apple Macintosh, IBM-PC and IBM compatibles, and IBM PS/2. The Macintosh is used most often for sophisticated word processing (desktop publishing) because of its excellent formula writing capabilities and graphic applications. IBM-PC compatibles still dominate the market and are highly competitive in terms of low prices, many suppliers, and the excellent availability of extra components and software. The IBM compatibles are used for routine word processing and administrative work. The PS/2 is far less

common and has certain advantages in that it can function under both DOS and IBM's new operating system, OS/2.

2.2 The PC networks

We have a number of different network experiments currently under way, and we will discuss a few. A network is any linking of different computers and terminals with other components so that all of the different components communicate. A local PC network is different from other nets in that the communication occurs at such high speeds across cables designed for PC networks. A standard network consists of servers, i.e., the computers supplied with software, printers, etc., and work stations. In addition, some networks have gateways that allow communication with other systems, for example, a mainframe. Bridges, i.e., the connection between different networks, can also be a part of a network, but at the present, there are no bridges between Statistics Sweden's networks.

A few of the networks are connected to IBM's System 36 mini-computer. There are three IBM System 36 networks with a total of 90 nodes, i.e., terminals, printers, etc. System 36 is used mainly for office automation and has not proved suitable for other types of work. This system will not be further developed by IBM, but will be replaced by AS/400. Statistics Sweden has no plans to buy an AS/400, but we do intend on expanding our system of PC networks.

There is a Macintosh network that uses AppleTalk communication software and consists of 20 nodes of which there are: one server, four printers and fifteen work stations. There is another AppleTalk net with four work stations and one laser printer.

There are a number of tests going on with IBM-PCs using IBM's Token Ring with IBM's PC-LAN as the network software. To be included in one of these networks, the PC must be equipped with a Token Ring adaptor. These computers are linked via an IBM cable system.

2.3 Software

Statistics Sweden has developed most of the software used on the mainframes for the production of statistics. These programs are standard and general in that they are used by most of Statistics Sweden's surveys to compile tables (TAB68), graphs (CueChart), calculate variances (SMED), and perform the other computer processing necessary to produce descriptive statistics. Like other national agencies, it is not routine for Statistics Sweden to analyze the data, although data analysis can be a part of consultation work and is done in some continuing surveys. We have also developed software for sample selection that is tailored for Swedish circumstances (our unique reliable registers, etc).

In the early 80s we compiled databases and developed a database management software, AXIS, for the documentation, presentation,

and distribution of tables and time series. AXIS allows the user to generate tables tailored for his/her specific needs. AXIS is different from other database managers in that the data are stored in such a way that unique tables and time series can be produced. AXIS is currently being used by Denmark's central statistical office for its database of municipalities and by the Economic Commission for Europe in Geneva.

On the other hand, the software used at Statistics Sweden on the PCs is commercially developed and available on the market. PCs, as stated earlier, are used primarily for word processing, administration, and some data analysis. Here, we have tried to limit the number of different software and promote a standard mix of software used at the agency. It was deemed uneconomical to have too much different and overlapping software because those responsible for installation, upkeep, and advice cannot be completely familiar with everything on the market. Those who choose to use an "exotic" software must take full responsibility for it and not request help from the agency's computer experts. On the other hand, help is readily available for the software endorsed by the computer science unit, e.g., SAS.

2.4 Policy

In the main, Statistics Sweden's EDP policy is that computer technology should eventually lead to a highly decentralized production of statistics. A survey's personnel will be responsible for every aspect of the survey from questionnaire design, data collection, to presentation and dissemination. This decentralized system requires that a survey statistician be something of a "jack of all trades" in that he/she is responsible for the survey from start to finish. The survey statistician is even expected to produce his/her own publications using sophisticated word processors (desk top publishing). This puts great demands on the individual statistician in that he/she must be fluent with a wide variety of software and must have a command of all aspects of survey practice. He/she is a self-sufficient unit supported by a well-chosen mix of generalized packages. These packages will be those developed both commercially and by Statistics Sweden.

3. Computer-assisted Data Collection

The development of computer-assisted telephone interviewing (CATI) has been in progress for almost two decades. Most users of CATI report good results. Nevertheless, CATI's presumed advantages regarding quality and costs compared to regular telephone interviewing seem to be based on logical assumptions rather than methodological studies (see Nicholls and Groves (1986) and Groves and Nicholls (1986)).

During recent years, Statistics Sweden has worked on developing an even more refined device for different kinds of data

collection. We are aiming for a general device for computer-assisted data collection (CADAC). The challenge consists of finding a device that could be used in a standard centralized CATI setting, for decentralized telephone interviews conducted from interviewers' homes, for face to face interviews conducted almost anywhere and for entry of data on prices, crop sizes, etc. To do this we need a hand-held portable computer.

The following summarizes our reasons for developing a hand-held computer.

1. Statistics Sweden needs a decentralized automated data collection system because our interviewers are scattered around the country. To use CATI only, whether centralized or decentralized, would put constraints on our automation resources.

2. Some of the survey work conducted at our agency is done on commission. To retain our share of the market, Statistics Sweden must be at the technological forefront.

3. Cost, quality, and timeliness are vital for the production of statistics. Any device that can save costs and improve quality should be tested.

The plans for the Swedish CADAC as they were outlined in 1984 were presented in Lyberg (1985). These plans included a set of specifications regarding the computer per se (weight, ergonomy, etc.), the manufacture of the computer, development of software, and testing. The set of requirements was presented to a number of potential manufacturers but only one was willing to attempt developing the computer. Most firms claimed that it would be too difficult to meet our weight requirement. (To allow for an ergonomically sound data collection it was decided that the weight should be less than one kilo and the screen should have space for 24x80 characters.)

The detailed set of requirements can be found in Danielsson and Mårstad (1983). Briefly, these requirements included (a) the ergonomics mentioned, (b) that the computer must be usable in all kinds of data collection environments, (c) texts, messages, and data must be transmitted over the telephone network implying a considerable data reduction, (d) storage would require a 512 KB memory, (e) there should be a battery capacity for one day's work without network connection, (f) response time should not exceed two seconds, (g) applications must be transferable from a central computer to the hand-helds within 48 hours, and (h) the system must be extendable and controllable.

3.1 Some tests

The project started in the early 80s by simply studying various American CATI systems. To gain CADAC experiences a first test of data entry on hand-helds was conducted in 1982. A computer called MICRONIC was used to collect prices in stores for the consumer price index. By and large the results of this first test were encouraging. The test had, of course, many shortcomings

with a small screen and low general storage capacity. But the interviewers liked the new technology and the technical evaluation of the test provided many valuable hints for future designs.

In 1984 a CATI test was conducted. The setting was standard for CATI facilities with display terminals connected to a mainframe. Since we know that such settings work, this study was primarily aimed at interviewer reactions. The overall interviewer reaction was favorable.

To test respondents' reactions to hand-held microcomputers, a study was conducted in 1985 using a portable computer, Data General/One, that was heavy but available on the market. No negative reactions from the respondents were noticed.

During 1986, a few hand-held prototypes were delivered. These prototypes had considerable shortcomings given the set of requirements that had been established. The most notable shortcoming was that the computer was too heavy. Nevertheless, a series of tests of the new equipment started late 1986. The surveys involved were the (Political) Party Preference Survey and the Labor Force Survey.

In 1987 a major change took place within the project. The Swedish manufacturer admitted that it was unable to proceed beyond the prototype series already developed. The firm claimed that it lacked the know-how and other resources necessary to meet all the technical requirements set up by Statistics Sweden. Although this declaration had been expected for some time, it caused considerable disappointment at Statistics Sweden. The software and technical solutions developed within the project are for the most part independent of the hardware, so we started to look for a computer already available on the market. Currently we are working with Toshiba 1200.

3.2 Testing Toshiba 1200 in the Labor Force Survey

In 1988 a test of the hand-held Toshiba 1200 was conducted in the Labor Force Survey. Thirty-five interviewers conducted altogether 821 interviews and the purpose of the study was to test the EDP system and software. This is the first step in a large-scale feasibility test where forthcoming steps will include methodological comparisons of quality and costs. Examples of factors under study in this first step included system reliability, screen visibility, ergonomic qualities, hand-held reliability, communication between the mainframe and the hand-held, and interviewer and respondent reactions. All of the 35 Toshiba 1200 computers that were used worked without major problems during the test. To keep track of problems, log-books were used both by the central staff and by the interviewers. These log-books contained information on the communication between interviewers and the central staff and descriptions of various problems that occurred during the interviews.

Even though the system worked fine, it can be improved in many ways. The problem with the weight of the hand-held is still with

us of course. The current weight is 5,2 kilos and is unacceptable for the kind of device we are aiming for. We also experienced problems with the batteries. In some cases, they did not last as expected. Two days of interviewer training seemed too short; this opinion is supported by the intense communication between interviewers and central staff during the first days of the test. There is strong evidence that the interviewers need more time to practice before entering actual survey work. The test also pointed out a number of shortcomings related to the software, shortcomings that we do not give an account of here.

The interviewers and a sample of respondents were asked about their reactions concerning this new technique. Respondents were more or less indifferent. Some interviewers were a bit reluctant towards the hand-held, though. Concerns were expressed regarding interviewing efficiency, physical stress, and the fact that the interview becomes strictly controlled. Most interviewers, however, thought that working with the hand-held was very interesting.

3.3 Where do we go from here?

The plans for manufacturing a new and extremely light-weight hand-held computer have currently come to a halt. The software development continues and can accommodate a recently set up CATI installation of eight work stations. There is a series of studies in progress that we hope will fill in the gaps in our knowledge about the feasibility of hand-helds. As for the hand-held computers themselves, we are reduced to the existing market. Depending on the development of existing hand-helds, different options can be identified. First, with the already available software the centralized CATI facility can be expanded to an optimal size. Second, if new light-weight computers appear on the market, Statistics Sweden can continue to refine its original concept. Third, if only heavy computers are available, centralized CATI can be extended to include decentralized CATI conducted from the interviewers' homes. Fourth, again, if only heavy computers are available our concept might include CATI and computer-assisted personal interviewing but excluding the forms of data entry where the computer has to be carried during the data collection. For the time being the third and fourth alternatives seem most likely.

4. Editing

Editing, i.e., the process designed to ensure that the collected data meet certain quality requirements, has a long tradition at Statistics Sweden. Editing is important not only for the elimination of erroneous values. It also serves as an instrument for quality control of the data collection. We can distinguish two main forms of editing, micro-editing and macro-editing. In micro-editing, inspection of individual observations is carried out following specified rules to find errors and to correct or at least adjust erroneous observations. In macro-editing, errors are found by inspecting groups of records.

With micro-editing, it is possible to combat some anticipated errors due to respondent negligence or misunderstanding. Surveys should be designed to minimize the risk of such errors which means that in well-designed surveys these error rates are fairly low and the errors can be handled by data entry editing. If the data entry computer is not powerful enough, existing generalized micro-editing software can be used. GRAN78, GRUS, and FELOC are examples of such systems. Yet, we know very little about the impact of editing on the quality. There are indications that micro-editing adds very little and in some cases the impact might even be detrimental. Thus, there is a need for a thorough evaluation of current micro-editing practices. When using micro-editing, there is a considerable risk of overestimating the data quality. What often happens is that once the date have passed a lot of micro-editing checks, neither the data nor the survey production system will be critically reviewed.

There are certain problems with software maintenance. The programs are old and have limitations. GRAN78 is written in ASSEMBLER and GRUS in APL. The competence at Statistics Sweden in these languages is rapidly decreasing and the general need to maintain this competence for other purposes is also decreasing. The current set of programs has severe limitations. Simultaneous correction of two or more files cannot be done, for instance, the case of a household file and an individual file in the same survey.

Statistics Sweden participates in the international Statistical Computing Project. Software developed in Canada (CAN-EDIT), Hungary (AERO), and Spain (DIA) have been studied but found unsuitable for Swedish purposes. Current studies examine the BLAISE system developed in the Netherlands and the Canadian portable GEI system.

Macro-editing is more rational, but its results have long run rather than immediate value. Errors detected by macro-editing can be prevented in future surveys, but not easily remedied in the present survey, for instance, the macro-editing might suggest the insertion of a new redundancy procedure.

The role of editing should be extended and editing should become another quality control instrument. Thus editing should be used primarily to evaluate the survey's quality and to provide quality information, even down to the single variable value level. That single values are as accurate as possible is especially important if the data are to be stored in a database. The role of editing should not be reduced to handling troublesome errors and tidying up data files.

Characteristic for early editing policies was that the editing should be designed so that small numbers of anomalies and missing values should not be permitted to disturb the subsequent data processing. One policy employed in the United States Census Bureau (see Ogus et al. (1967)) was that not more than five percent of the data should be adjusted (usually by imputation)

in the editing process. That sophisticated imputation methods are nowadays often used as a means of dealing with nonresponse and editing is seen as a quality control instrument highlights the need for a new editing policy.

In some applications at Statistics Sweden, the editing is extensive with lots of checks; this can result in a large number of error indications. Since only a fraction of the indications turn out to be real errors, those responsible for surveys using extensive editing should consider allocating more resources to preventive measures, for instance, pretests, improved questionnaire development, and interviewer training. There is a risk of suboptimization as a result of what we call "the control device approach." The very existence of a control device seems to justify its use. Using the available technology, whether it is necessary or not, is common in nonsampling error work where some instruments have been developed (like editing and key punching controls) and other more needed instruments have yet to be developed (like instruments to be used to control certain measurement errors). Nevertheless, not everything that can be checked should be checked. As stated by the United Nations (1982) "The gradual expansion of computer editing and imputation procedures has caused some users and producers of data to be uneasy. After all, the purpose of collecting data is to discover or reaffirm some elements of truth about the population. To make extensive changes in the collected data prior to making them available for analysis is to violate a basic principle in data collection that the integrity of the data should be respected."

So, editing must be carefully weighed against other control measures. The present attitude at Statistics Sweden is that too many resources are spent on editing. Editing should be considered a survey methodology problem rather than an EDP problem. Without reducing data quality, editing costs can be cut by a careful choice of editing checks, conducting most of the editing at the data capture phase, and using macro-editing rather than micro-editing.

A new user friendly system founded on these principles is now underway. A first outline of a new editing system has been worked out and distributed to the surveys. The reactions to these plans vary. Editing has always been a controversial issue at Statistics Sweden. It has been decided, however, to proceed with the development work. The design of a new editing system should take into account other developmental work conducted at the agency, including editing as a part of the computer-assisted data collection system and integration of data capture and editing.

5. Automated Coding

Coding is a major operation in the Swedish Census of Population and Housing and a number of other surveys conducted at Statistics

Sweden. Coding transforms verbal information to code numbers to facilitate data processing. When carried out manually, this process is susceptible to errors. A number of studies show that the error rate, when coding complex variables like occupation and industry, can be extensive. An evaluation study of the coding in the 1970 Swedish census gave the following results.

Table 1. Estimated error rates in coding economic activity in the 1970 Swedish census of population

Variable	Code	Percent error rate (total population)
1. Relationship to head of household	1-digit	4.3
2. Type of activity	1-digit	4.7
3. Occupation	3-digit	13.5
4. Status	1-digit	3.7
5. Industry	4-digit	9.9
6. Place of work	1-digit	8.9
7. Type of conveyance to place of work	1-digit	11.5
8. Number of hours at work	1-digit	4.4

The table shows that the multi-digit variables are difficult to code. But the one-digit variables, a priori considered easily coded, are also erroneously coded relatively often.

When preparing the data processing of the 1975 Swedish census, new quality control procedures and administrative changes were considered. This resulted in a tighter verification program and the error rates decreased for every variable. An evaluation study provided the following results.

Table 2. Estimated error rates in coding economic activity in the 1975 Swedish census

Variable	Code	Percent error rate (total population)
1	1-digit	0.6
2	1-digit	0.6
3	3-digit	7.8
4	1-digit	0.5
5	4-digit	3.5
6	1-digit	1.0
7	1-digit	0.5

Extensive reviews of Swedish studies of error rates in coding are given in Lyberg (1981, 1983).

During the 70s, it appeared important to analyze manual coding in detail to consider the possibilities of computerization. It had been shown that the efficiency of manual coding could be improved by using better manual control techniques but that improvement was considered insufficient. An automation was called for. At the time automated coding had been considered by the U.S. Census Bureau for some time (see O'Reagan (1972)) but the method had not been applied in real surveys.

In automated coding we distinguish four operations. First, a computer-stored dictionary must be constructed. Second, the verbal descriptions are entered into the computer. Third, the verbal descriptions are matched with the dictionary and codes are assigned when matches occur. Fourth, regular evaluation of the automated process must be performed as a check on cost and quality.

Several methodological problems are inherent in automated coding. The most interesting ones are those associated with the dictionary and the matching algorithms. The dictionary itself is best constructed by a computer program. The input, however, must consist of manually expert-coded descriptions preferably selected from the first part of the survey to be coded. When constructing the dictionary two lists of expert-coded descriptions are used. List No. 1 consists of coded descriptions by code number. List No. 2 is the same file where the coded descriptions are coded alphabetically. From list No. 1 the most frequent descriptions are selected to be included in the dictionary. List No. 2 is used to check that these frequent descriptions have been assigned the same code number. It is rather common that a specific verbal description can be associated with different code numbers and list No. 2 helps us identify these ambiguous descriptions so they can be deleted from the dictionary. The matching algorithm is straight-forward. Only exact matching is used which means that a certain proportion of verbal descriptions is referred to manual coding when automated coding is applied. Research at the

U.S. Census Bureau has focused on more complicated matching algorithms using weighting. That work is documented in, for instance, Appel and Scopp (1985) and Knaus (1987). Statistics Sweden has not used or even experimented with these algorithms.

Applications of automated coding at Statistics Sweden include the coding of goods in the 1978, 1985, and 1988 Household Expenditure Surveys (HES). The design of the 1978 HES included continuous delivery of diaries kept by the respondents during the survey year, resulting in the material being processed in cycles. In 1978 the initial dictionary was based on an experimental coding conducted the year before. During the survey year the size of the dictionary increased from 1 459 to 4 230 descriptions as a result of 16 revisions. The degree of success can be measured by the proportion of descriptions coded by the computer (coding degree). Table 3 gives the dictionary sizes and the coding degree for the 33 cycles.

Table 3. Dictionary size and coding degree for the 33 cycles in the 1978 HES

Dictionary version	Cycle	Number of dictionary descriptions	Coding degree (%)
1	1	1459	56
2	2	1554	63
3	3	1760	67
4	4	2228	66
5	5,6,7	2464	68,68,63
6	8	1632	64
7	9	1990	53
8	10,11	2451	69,66
9	12	2866	61
10	13	3065	68
11	14	3613	58
12	15,16	3752	72,73
13	17,18	38323	39,70
14	19,20	4011	65,73
15	21,22	4229	51,72
16	23,24,25,26 27,28,29,30	4230	64,67,62,67 72,65,67,50
17	31,32,33	4230	65,39,67

The coding degree over all cycles was 65%. Two different diary forms were used which explains the sharp decreases in coding degree from time to time. One of the forms was much more suitable for automated coding than the other.

In the 1985 Household Expenditure Survey essentially the same procedures were used. We have data on the first 17 cycles in that survey as shown in Table 4.

Table 4. Dictionary size and coding degree for the first 17 cycles in the 1985 HES

Dictionary version	Cycle	Number of dictionary descriptions	Coding degree (%)
1	1,2	1985	81,78
2	3	2029	82
3	4,5	2063	82,81
4	6	2126	82
5	7	2156	83
6	8	2176	82
7	9	2207	81
8	10	2228	83
9	11,12,13	2272	83,83,82
10	14,15,16	2370	83,80,81
11	17	2446	83

The 1988 HES is still in process but we know that the coding degree is fairly stable around 80%.

A few observations regarding the automated coding of Swedish Household Expenditure surveys are the following. (1) Special evaluation studies show that the error rate for the part that is automatically coded is approximately 1%. (2) Automated coding is slightly cheaper, 2-5%, than manual coding. (3) The automation has created a better general coding environment in that the remaining manual coding has become more interesting to the coders and the administration has been reduced. (4) It does not seem worthwhile to make extensive revisions after a specific point. A stable coding degree is obtained soon and not even extensive dictionary updating can increase the coding degree substantially. The coding degree does not increase notably when the dictionary increases provided that the initial dictionary version has been constructed according to the methods we have described. (5) The computer coding is very inexpensive. Automated coding of tens of thousands of descriptions costs less than 50 dollars.

Another important application is the automated coding of occupation in the Swedish Census of Population and Housing. Verbal descriptions of occupation are keyed and edited. In the editing process, special signs and prefixes are removed and the remaining parts of the description are brought into one sequence. Each record is matched against a file containing the economically active population. As a result, we get a matched file that contains the edited verbal description and auxiliary information for each record. The edited file is matched against a computer-stored dictionary and if an occupation description is found in the dictionary a code number is assigned. Records that contain descriptions not found in the dictionary are manually coded.

The manual coding is computer-assisted and carried out on display consoles and in two steps. The initial manual coding is done without access to the questionnaires and is a fast operation. Records that cannot be coded are left "empty" and taken care of in the second manual coding where the coder has access to the questionnaires.

The dictionary consists of two parts, a primary dictionary, PLEX, and a secondary, SLEX. PLEX contains unequivocal descriptions and is scanned first. SLEX contains discriminating word strings that fit several different input descriptions. As a consequence SLEX is not as accurate as PLEX and is used only if PLEX fails to code.

In the 1980 Swedish census, the size of PLEX increased from about 4 000 records to more than 11 000 during the course of the census. The SLEX size was kept at approximately 500. The coding degree for the entire census was 71.5%, roughly 68% by PLEX and 3% by SLEX. The coding was conducted county by county and the coding degree varied between counties from 67% to 77%. As was the case with HES, the matching cost is negligible. In the following example the descriptions for one county with 314 529 economically active individuals were matched against a PLEX containing 10 291 descriptions and a SLEX containing 513 word strings. The total coding degree was 74.7% and the cost was approximately 50 dollars.

In the 1985 Swedish census the coding system was the same as in 1980. The only main difference was that the size of PLEX was increased to 20 000 and the size of SLEX to 1 900. The coding degree for the entire census was 73.7%, roughly 69% by PLEX and 5% by SLEX. The coding degree between counties varied between 72% and 75.8%.

The following observations regarding automated coding of occupation in the censuses can be mentioned. (1) The cost of automated coding is approximately 10% lower than that of manual coding. (2) The coding becomes faster. (3) Because of the automation, the manual coding workload is lessened and it is possible to use only the regular staff rather than a great number of temporarily hired coders. For a few years, automated coding of occupation has also been used in the Swedish Labor Survey. In this survey, we use a dictionary consisting of slightly less than 5 000 descriptions. The coding degree is approximately 70% with very small monthly variations.

Automated coding is used in some other surveys as well. The most notable example is the coding of book loans in a survey sponsored by the Swedish Authors' Fund. Automated coding has been used for ten years and during that period, the dictionary size has grown from 7 000 to over 65 000 authors and book titles. The current coding degree is approximately 75% and the system is highly profitable compared to the old manual one.

6. Applications of New Techniques for Regional and Local Planning

Area-related statistics on urban expansions, recreational areas, agriculture, and land-use are of great interest in Sweden. Information is collected to provide statistics for urban and regional planning and for managing natural resources. This kind of data collection would be virtually impossible without access to modern techniques. Maps, aerial photos, and to some extent satellite information are used as data sources for land-use statistics and geographic variables are improved by the use of coordinates.

Statistics Sweden has experimented with remote sensing (aerial photos and satellite data) for a few years. The main conclusions from these experiments are that remote sensing, for the time being, cannot meet reasonable precision requirements and that it is too expensive. Traditional survey methods like sampling and collecting data via mail questionnaires are more efficient. There are, however, specific remote sensing projects that might be worth working on. For instance, crop yield estimation is carried out by means of a staff of 400. Each year these persons harvest selected one square meter crop areas scattered over approximately 50 000 fields. This is an expensive procedure and it is important to determine whether remote sensing can be used to get auxiliary information. Another example is the Swedish system for insuring crop damages. It is quite clear that this system cannot be entirely based on satellite data, but aerial photographs could be used as a complement to the system. For example, aerial photographs have been used to document the extent of flooding. A similar application is the use of aerial photographs to demarcate damaged farm land for farmers who claim more damage than the area classification defines. Remote sensing could also be used to get summary information on crop characteristics and thus make a manual follow-up more directed and efficient.

A lot of the land-use statistics are based on registers and satellite data cannot compete with these registers due to costs. Registers, however, do not provide information on green areas and waters in densely populated areas. This kind of information could be collected by satellites as could information on expansion of densely populated areas and changes in areas of common access. (Common access is a privilege granted by Swedish law that permits individuals to enter privately owned lands except for cultivated fields and private grounds near dwellings.)

A general evaluation shows that remote sensing is more useful in collecting land-use statistics than crop-related statistics. Most of the work on remote sensing is still developmental. The skies in Sweden are clear only every fifth day on the average, and conventional survey systems used at Statistics Sweden are efficient enough to justify a relatively slow progress in the use of remote sensing (see Hall-Könyves (1988) and Justusson (1984)).

New technology, in a broad sense, is vital for producing statistics for regional and local planning (see Ansén et al. (1988)). For instance, statistics on land-use in urban areas and their immediate vicinities have been prepared by using map-

aided plotting from aerial photographs (see Matérn (1986)). In sampled urban areas, land-use is determined via selective plotting from aerial photographs. Data from existing urban area maps and economic maps supplement the information obtained from the photos. The plotting is done on light tables from prints of black-and-white aerial photos. A transparent plastic film with an etched 1-cm grid is placed over the photo and old urban-area demarcation lines are drawn on the film. Outside these demarcation lines, new boundaries are drafted showing the outer limit of the zone of accessibility (real distances of 500 or 1 000 meters depending on the population size). Points are then plotted corresponding to various real distances of 200 meters and less. The type of land-use, according to a special classification system, is coded directly on the plastic film. This is an example of what is called a cartographic survey.

Another example of new technology is the use of coordinates. All traditional real-property registers are currently being replaced by a computerized system, and these will be finished within a few years. When setting up the new registers, coordinates for each real-estate unit are added. Then data linked to real property can be located for arbitrarily delineated areas, and new possibilities open up when computer-based spatial analysis and cartographics are available. In the future, it will be possible to make more extensive use of geographical characteristics and this is important since density, potentials, spatial covariation, and other geographic variables are so useful to statisticians and planners. Various agencies in Sweden collaborate to produce computer-based statistical maps (dot, grid, isarithmic) and scales.

7. Issues Related to Privacy and Confidentiality

During the last two decades, Statistics Sweden has devoted a great deal of work to protecting the privacy and confidentiality of its data collection and storage. In 1973 Sweden became a pioneer in data protection by passing a national Data Act for the protection of privacy and confidentiality in both the public and private sectors. The main incentive for the law was the intense public debate about the data collection procedures and the contents of the 1970 Census of Population and Housing. The 1970 census employed old-fashioned data collection procedures involving, for instance, janitors as census form collectors in apartment buildings. Also a new section on educational achievement was introduced. In some countries, income is a sensitive survey variable, in Sweden education is. The 1965 Census of Population and Housing was a smooth operation where the agency received just a few telephone calls during the data collection period. In 1970 the switchboard was blocked for three weeks, and during these three weeks, the central census staff did nothing but talk to concerned and angry respondents. Despite that the vast majority of households filled in their forms without reminders or follow-ups, Statistics Sweden felt that it must take immediate actions to remedy these new problems.

As a result, Statistics Sweden started to develop a number of protection measures on its own or together with other agencies and authorities. Also, a new agency, the Data Inspection Board, was established to supervise agencies and organizations in their handling of data. Over the years, Statistics Sweden's data protection activities have not diminished. Public concern was expressed about the 1980 census. In 1983, a commission appointed by Statistics Sweden proposed that the 1985 census should be based on data from existing records rather than on data collected directly from households. The mass media, various organizations, and some private citizens expressed their concern about this proposal and eventually the government decided that the data collection should not be changed in the 1985 census. As soon as the data collection for the 1985 census was completed, there was a new outcry. The National Tax Board was, according to the population census law, allowed access to the address data in the census forms as a check of the board's own registers. Before filling in the census forms, however, the households had been informed that census data would be used for statistical purposes only. Since the board's address check was part of a larger plan to prosecute tax evaders, the permission was vehemently criticized.

In 1986, a Swedish sociological research project, a longitudinal study called Metropolitan, started an inflamed debate. This project was led by a research group at Stockholm University and Statistics Sweden was only marginally involved. Because most of the data collected came from administrative records and many of the 15 000 individuals were unaware of their participation, the debate concentrated on privacy issues. The debate was intense and continued for six months with daily media coverage. This debate, like the previous debates on privacy issues, decreased the public's willingness to participate in Statistics Sweden's surveys. The nonresponse rate in the Swedish Labor Force Survey almost doubled in the first few months of the debate. Only through extensive efforts could Statistics Sweden even moderately reduce the nonresponse rates and to this day some of our surveys have not fully recovered.

The computer revolution has posed certain threats to an individual's privacy in that our capabilities to collect, copy, store, merge, and retrieve information about individuals have increased dramatically. On the other hand, it is the same technology that allows us to protect the individual from potentially abusive data surveillance. Dalenius (1988) provides a review of measures to protect the integrity of survey data. These measures are not only technical, like de-identification, suppression, rounding, etc., but are legal and administrative as well.

Today, Statistics Sweden routinely practices a number of protective measures. First, it is Statistics Sweden's policy to de-identify all materials that are of limited value to researchers or where no future needs for linkages can be predicted. De-identification is considered the surest and most practical safeguard. It can be done in connection with the data collection, when the data processing is finished, or after some prespecified

time. The policy implies that registers created by record linkage are de-identified and sometimes even destroyed after being used. Some materials like those on political party preferences are de-identified despite the limitations that this puts on future research. This is done for privacy reasons only and this part of the policy shall be applied even more frequently in the future. It should be pointed out that de-identification is not 100% sure. On the other hand, backwards identification is difficult to perform since it demands considerable resources and detailed information about the individuals.

Other measures are administrative, for instance, the individual's name and address can be separated from the individual's personal identification number. Access to information is also limited by security measures and all visitors must identify themselves.

A number of laws and regulations are at work too. The Data Inspection Board must give permission, or in simple cases, a general licence, before new registers can be created. Every employee must sign a secrecy declaration and the punishment for violating this contract is severe indeed. Protected information cannot be used by witnesses in trials unless the offense is so severe that the court chooses to overrule that rule. So far, courts in Sweden have never exercised this right. Statistics Sweden has access systems that limit the number of persons allowed to work with the different registers. The newly installed access system, Top Secret, allows access to only certain individuals and only at prespecified terminals. Theoretically, Top Secret could be circumvented by reading data directly from the disks or by bugging the computer-terminal communication devices. This, however, would be an extremely awkward method resulting in virtually incomprehensible sequences of number. At present, we do not feel that the risk posed by bugging is serious enough to justify protective measures.

The total de-identification of material makes future research, like longitudinal studies and record linkages impossible. A technique that allows de-identification while still permitting record linkage is needed. One such technique is the one-way encryption of personal identification numbers. With all identification numbers encrypted, sensitive data can be processed without the risk of an individual being linked to his or her data. In one-way encryption, the encrypted number replaces the identification number and yields virtually no information at all. This assumes that there is no other feasible means of identification in the register.

The first step in encrypting personal identification numbers is that these numbers are mapped to a set of encrypted numbers using a complicated algorithm. This algorithm is inversely unambiguous which means that two different identification numbers will always be mapped to two different encrypted numbers. The algorithm should be such that it is virtually impossible to decipher the encrypted number. According to the data law, the encrypted register is no longer a register of individuals. Nevertheless, the encrypted register is still kept confidential because of

the small risk posed by backwards or other technically difficult identification.

After encryption, the unencrypted file is stored at the national archives and no remnants are left at Statistics Sweden. Even the medium that contained the unencrypted file is demagnetized.

The one-way encryption algorithm can be created in various ways. The route chosen by Statistics Sweden is to use a conventional algorithm that can code and decode with equal speed. Both the algorithm and its inverse are top security secrets. The final encryption program is arrived at after several steps and exists only as a loading module protected by the new authorization system. Details about the method of one-way encryption are found in Rapaport and Block (1986). Recently the Data Inspection Board gave permission to Statistics Sweden to experiment with one-way encryption during a three year period in two surveys.

Statistics Sweden has had a research program on statistics and privacy for many years and the agency has cooperated with many other agencies and organizations in Sweden and abroad. Two international conferences have been arranged on privacy and confidentiality issues, one in 1978 and one in 1987. The latter is documented in Statistics Sweden (1988).

8. Dissemination

Statistics Sweden has two main principles for the dissemination of its statistics. The first is that statistics should be easily accessible for the user and suit the user's capacity to utilize the statistics. The second is that the average citizen should be kept well informed about the important developments in our society. To reach the general public, Statistics Sweden sends copies of publications to libraries and sends press releases to be disseminated through the mass media. We do not offer any financial remuneration for the dissemination of the agency's statistics.

To the contrary, we often require that users pay for publications and statistics on other media. This provides some measure of the demand for these statistics and helps cover the cost of dissemination. It is not our policy to demand the highest price possible, but to find the optimal level of dissemination. Electronic dissemination is costly, since the research and development costs must be covered and the same statistics are usually available in less expensive forms, for example, in publications. A certain degree of dissemination is done free of charge to, for example, municipalities and other governmental agencies and organs. Those who participate in surveys are usually given a copy of the resulting publication so they know what they have participated in.

8.1 Publications

Statistics Sweden produces 700 regular publications per year that sum up to 600 000 single issues. A steadily increasing

percentage of these publications are being produced by sophisticated word processing techniques. It is the official policy that the production of statistics should be as decentralized as possible, and the publication step is no exception. Another goal is to replace a number of publications with electronic media.

8.2 Electronic media

The electronic media that Statistics Sweden uses to make statistics available are magnetic tapes, floppy disks, compact disks with read only memory (CD ROM), videotex, and on-line databases. Each medium has its advantages and disadvantages and can be considered a complement to or a replacement for published reports.

8.3 Magnetic tapes

For a moderate fee, Statistics Sweden releases data on magnetic tapes. Some users receive tapes on a monthly basis. There is not, however, so much demand for data on magnetic tapes, so this medium cannot be considered a significant dissemination channel.

8.4 Floppy disks

We have disseminated data on diskettes for the past few years, but only on a modest scale. As of yet, there are not any principles or standards for making data available on diskettes. Nor is there any software tailored for handling diskettes from Statistics Sweden.

8.5 Compact disks with read only memory

Data dissemination via CD ROM is still in the experimental stages. The goal is to make large data sets available on CDs. At the present, a few Swedish libraries can accommodate CDs, and the Swedish school system has expressed interest in large data sets stored on CDs for dissemination through the school system.

8.6 Videotex

Videotex can be described as a system of statistical information, i.e., tables and diagrams with explanatory notes and analyses set up in such a way that the user can browse through the system using menus to retrieve information of interest. It is purely visual, even though the "video" might lead one to expect sound. During the past three years, we have conducted a large scale test of videotex. Swedish schools showed a great deal of interest and a system is in use, but on a lesser scale than that used during the test. The current system has about 50 different videotex pages.

8.7 Databases

Since the early 80s, Statistics Sweden has used statistical

databases as an alternative to publication. Of the electronic dissemination options offered, the databases, the Time Series Data Base (TSDB) and the Regional Statistics Data Base (RSDB) are the most used. These databases employ the database manager AXIS and the user communicates on-line with Statistics Sweden's computer to generate user-tailored tables. Here it should be pointed out that AXIS is more flexible than other database managers. The data are stored in a way that allows for the generation of user-designed tables. One disadvantage of the AXIS system is that it can never be used as a full-text data base. Because the user designs the tables, tables are not stored in the database and thus no commentary more extensive than footnotes explaining variables etc., can be stored. This characteristic means that AXIS can never fully replace the publication of Statistics Sweden's regular report series, Statistical Reports, since only a full-text database can include the commentary and analysis that is characteristic for a Statistical Report.

Statistics Sweden's current database system was introduced in 1982. After three years, there were 110 on-line subscribers. After another three years, in 1988, the number had doubled and is now 220. The number of actual users is estimated at around 500. The number of stored matrices is approximately 1200 which corresponds to about 150 meters of published material in bookshelves. The databases managed by AXIS contain roughly 30% of the statistics published in reports. Some subject matter areas are not represented in the databases at all.

The computer time that database subscribers buy from Statistics Sweden to process the data are decreasing drastically. Many subscribers have their own PCs from which they retrieve the desired data from the database and on which they process and analyze the data.

The databases are expensive to run and the approximate cost is 10 million SEK per year (2 million Canadian dollars). Income from the users is about 1.5 million per year. Nevertheless, it is our goal that the income from the databases will eventually cover the costs. Because the databases are a new medium, they are expected to eventually become financially self-supporting. This financial demand is not put on the publications. One proposal for diminishing the gap between the income and cost of the databases is to drastically increase the subscriber fees. It is our long run goal that as society uses more electronics, more and more of what is now published in reports will be transferred to the databases and the databases will receive the financial support that the publications now receive.

8.8 Full-text databases

We have not yet used full-text databases as a dissemination medium. Nevertheless, there is a great deal of material that is well-suited for dissemination via full-text databases. The best suited material is the statistics whose value depend on the speed of dissemination. One example is the press releases.

Another example is the statistical reports, at least some of them. Each year, we publish approximately 400 press releases totalling 220 000 single issues. The financial saving would be enormous if press releases were disseminated via on-line full-text databases.

A word of caution about dissemination through statistical databases and other electronic media is warranted. A great deal of information is included in a statistical publication that should be included on the electronic media. Publications always include some type of data quality statements, the name of a contact person who can answer questions, and in some cases even technical information about how the data was collected and processed. There is a risk that any figures delivered in an attractive table or other diagram are taken as unquestionable truths. There is also a misconception that information that is not quantitative is not really information at all. We who work at national agencies are aware that there are reliable data and unreliable data. But how is the public, who is exposed to increasing amounts of data, to judge data quality. As electronic media are developed and their availability increases, the problem of data quality statements must be solved.

9. Future Developments

Two main developments can be recognized in survey work at Statistics Sweden. First, computers have taken over entire operations or parts of operations. In this paper we have reviewed the automation of coding and interviewing. The degree of automation varies from almost complete automation to various forms of computer-assistance. Second, there are attempts to integrate several operations into fewer. From one perspective, these developments are problem-free. The agency can cut costs and personnel and there is room to do new things and more things. From another perspective, however, technological advances create problems. Automated systems must protect confidentiality, be simple, fully tested, and highly reliable. In other words, automation must be fool-proof. Integration means that often one person takes over the duties formerly performed by several. Even though a new technology can simplify certain jobs, the technology still places new demands on the staff.

Recent developments and trends towards decentralization have led to what we might call a PC revolution. This calls for a totally new approach to educating and training of our staff. To this end Statistics Sweden has started an extensive continuing education program for its employees. The program will provide courses in many different fields and the education is offered to people outside the agency as well. One set of courses is intended for users of statistics while another is directed towards producers. To our knowledge this program, ranging from two hour mini-courses to full Ph.D. programs with its own staff of teachers including two professors has no counterpart among other statistical agencies in the world.

It is imperative that quality and cost considerations as well

as human factors are allowed to dominate the technology.
For almost a decade Statistics Sweden has suffered from severe budget constraints and there has been a tendency that personnel have been replaced by technology. During this adjustment period emphasis has been on producing the same amount of statistics or even more with smaller staffs and less than adequate funds. Rather little has been done when it comes to methodological studies like evaluation, pretests, and experiments not to mention continuous quality control. Much more of that was done in the early 70s. It seems, however, as if priorities will change again. All statistical programs are currently being scrutinized which will lead to redesigns and improved procedures. Statistics Sweden's personnel have to a large extent adjusted to the extensive use of PCs which should give room for increased work on quality issues.

10. References

Ansén, H., Hallén, S.-Å., and Ylander, H.(1988): Statistics for Regional and Local Planning in Sweden. Journal of Official Statistics, Vol.4, No.1, pp.35-46.

Appel, M.V. and Scopp, T.(1985): Automated Industry and Occupation Coding. Paper presented to the Census Advisory Committee of the American Statistical Association and on Population Statistics at the Joint Advisory Committee Meeting, April25, Rosslyn, Virginia.

Dalenius, T.(1988): Controlling Invasion of Privacy in Surveys. Handbook, Statistical Research Unit, Statistics Sweden.

Danielsson, L. and Mårstad, P.-A.(1983): Datorstödda insamlingsmetoder (DATI). Kravspecifikation för ett nytt system för fältdatainsamling i statistikproduktion. Memo, Statistics Sweden (In Swedish).

Granquist, L.(1984): On the Role of Editing. Statistical Review, No.2, pp.105-118.

Granquist, L.(1987): On the Need for Generalized Numeric and Imputation Systems. Report prepared for Seminar on Statistical Methodology, Conference of European Statisticians, Economic Commission for Europe.

Groves, R.M. and Nicholls II, W.L.(1986): The Status of Computer-Assisted Telephone Interviewing: Part II-Data Quality Issues. Journal of Official Statistics, Vol.2, No.2, pp.117-134.

Hall-Könyves, K.(1988): Remote Sensing of Cultivated Lands in the South of Sweden. PhD thesis, No.102, Lund University Press, Lund, Sweden.

Justusson, B.(1984): Fjärranalys- något för SCB? Memo, Statistics Sweden(In Swedish).

Knaus, R.(1987): Methods and Problems in Coding Natural Language Survey Data. Journal of Official Statistics, Vol.3, No.1, pp.45-67.

Lindgren, A.(1988 a): Microcomputers in Statistical Agencies. Memo, Statistics Sweden.

Lindgren, A.(1988 b): Lägesrapport från projektet runt lokala nätverk. Memo, Statistics Sweden (In Swedish).

Lyberg, L.(1981): Control of the Coding Operation in Statistical Investigations- Some Contributions. Ph.D. thesis, Urval, No.13, Statistics Sweden.

Lyberg, L.(1983): The Development of Procedures for Industry and Occupation Coding at Statistics Sweden. Statistical Review, pp.139-156.

Lyberg, L.(1985): Plans for Computer-Assisted Data Collection at Statistics Sweden. Bulletin of the International Statistical Institute, 18.2., Amsterdam.

Matérn, B.(1986): Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations, 2nd edition. Lecture Notes in Statistics, Springer Verlag, Berlin.

Nicholls II, W.L. and Groves, R.M.(1986): The Status of Computer-Assisted Telephone Interviewing: Part I-Introduction and Impact on Cost and Timeliness of Survey Data. Journal of Official Statistics, Vol.2, No.2, pp.93-116.

Ogus,J., Pritzker, L., and Hansen,M.H.(1967): Computer Editing Methods- Some Applications and Results. Bulletin of the International Statistical Institute, pp.442-466.

O'Reagan, R.T.(1972): Computer-Assigned Codes From Verbal Descriptions. Communications of the ACM, Vol.15, No.6, pp.455-459.

Rapaport, E. and Block, H.(1986): One-way Encryption of Personal Identity Numbers: A report on and Suggestions for Protective Measures. Memo, Statistics Sweden.

Statistics Sweden (1988): Statistics and Privacy. Report from a Conference in Stockholm, Sweden, June 24-26, 1987.

United Nations (1982): Survey Data Processing: A Review of Issues and Procedures. National Household Survey Capability Programme. United Nations.

HYPERQUEST: A NEW METHOD FOR INTERVIEWING USING LAPTOP COMPUTERS

M. A. JARO,

Systems Automation Corporation

HyperQuest: A NEW METHOD FOR INTERVIEWING USING LAPTOP COMPUTERS

Matthew A. Jaro
System Automation Corporation

ABSTRACT

This paper describes a new system that combines the interviewing process with a hyper-relational database. The interviewing engine has a full interviewing capability (multiple choice(s) from fixed or variable lists, long text responses, backing up to any previous question, help files, editing, loops, complex question paths, etc.). Instead of responses being written to a flat file, they are written to a hyper-relational database. The nature of this database is such that all data are automatically related down to the field level. A special query processor can be used to "browse" the information in a structured manner, or to make specific queries. The query processor is an object-oriented system that permits any user (with no assumed knowledge of the database contents or structure) to perform complex queries and to generate reports. Statistical information can be obtained about the responses. This concept is much different from that used in standard database systems where ad-hoc queries are either slow, complex to learn (SQL) or impossible without detailed knowledge of the database entity-relationships.

1. INTRODUCTION

Computer Assisted Data Collection (CADAC) and Computer Assisted Personal Interviewing (CAPI) have become increasingly important with the advent of "lap-top" portable personal computers. The current hardware technology now permits an interviewer to carry a light-weight battery-operated computer to the interview site and to conduct an interview. Better screen resolution and contrast, inexpensive hard disks, higher capacity floppy disks, large internal memory sizes, battery operation and relatively low price have all contributed toward making CAPI a feasible and technologically attractive technique.

The benefits of CAPI as a technique are obvious: online editing for completeness and consistency reduces the likelihood of followup visits due to edit failures in a mainframe-oriented processing environment; since the computer asks the appropriate questions, complex skip patterns and alternative paths can be introduced into the questionnaire without complicating training; printing costs and physical layout problems associated with hard-copy questionnaires can be eliminated; for multiple visit surveys, the responses from previous visits can be stored, and editing can be provided to check on longitudinal consistency; data entry or optical scanning costs are reduced since the results are machine-readable; write-in responses can be used more freely, since the interviewers enter the responses themselves, instead of data entry clerks trying to interpret the interviewer writing.

It is unfortunate that progress in software has not kept up with these hardware advances. Current interviewing systems are fairly difficult to use and create "flat" sequential files as output. The screens are rather confusing, operations are complex and querying results is generally impossible or at best inflexible.

To overcome these problems, we have developed the HyperQuest system. It combines an interviewing module, hyper-relational data base and report generating system into one package that can be loaded on a portable computer. This paper describes some of the aspects of this system that make it unique.

2. THE INTERVIEWING MODULE

2.1 The Interviewer Interface

A major requirement for an interviewing system is ease of use. Cluttered screens with multiple questions and non related information are difficult for interviewers to use. It must always be remembered that interviewers generally have no computer experience and are often intimidated by the idea of computers. For this reason, we have chosen a screen format that only contains a question box, an answer box and a status line. This format remains invariant through all questions so that a consistent, clutter-free interface is always presented to the interviewer.

Multiple choice responses are selected by means of moving the cursor. Help screens can be written for each question, so that if the interviewer does not remember the rules associated with a response, the F1 function key can be pressed to activate one or more help screens. The help screens can be easily "painted" by means of a special help screen editor.

The interviewer can interrupt the interview by pressing a single function key. The results of the system are automatically saved. When the interviewer begins again, the system is automatically restored, and the interview continues where it was left off.

Complex interviews often require that "flash cards" be shown to the respondent. These "cards" can be created by "painting" them on the screen with the help editor and flashed automatically where required.

2.2 Interviewer Movement

It is important that the interviewer be able to visit any previously asked question and possibly alter a response. Such alterations can affect the path of the interview from that point. The Hyperquest interviewing module automatically provides the capability of backing up one question, or backing up to any previous question (from a list displayed on the screen). If a question occurs in a loop, then the interviewer can back up to any occurrence of the question desired. Whenever backwards motion causes a previous question and response to be displayed, there are a number of options: 1) the interviewer can go forward to where he or she was before going back; 2) the interviewer can alter the response and go forward automatically to either the last question asked, or if the path is changed by the new response, to the first question that must be asked; 3) the interviewer can scroll forward through the path one question at a time.

William L. Nicholls and Carol House ["Designing Questionnaires for Computer-Assisted Interviewing: A Focus on Program Correctness"; U.S. Bureau of the Census, Third Annual Research Conference, Proceedings, 1987], discuss the issues involved in backward movement and subsequent forward movement. We have used those guidelines in designing the HyperQuest Interviewing module. In addition, that paper presents a number of principles that should be followed in designing a correct survey instrument.

In practice, the interviewer module keeps an internal stack of questions asked. Backward motion causes the database to be rolled back to the time of the question to be displayed and all intervening questions are placed on a pending stack. Responses are taken from the pending stack when appropriate. This feature prevents re-entering questions that have already been answered.

2.3 The Script Programming Language

The interviewing process is driven by a powerful fourth generation language, called SPL (the Script Programming Language). This language contains a number of special features that make the coding of very complex interviews possible. The following is a list of some of the capabilities:

> Special list variables and list-processing routines are available to handle menu choices and queries that can result in more than one response (multiple choice questions).
>
> Programmable help facilities are available to assist the end user in completing an interview.
>
> Interviews can be interrupted at any time and automatically restarted at the next question.

Question branching can be based on database contents, so that intelligent questions can be asked, consistency of answers checked and questions can be dynamically composed.

Powerful editing functions are available so that users can edit long entries. A number of text operations are available to manipulate user responses.

Menu choices can be read from ordinary ASCII files so that scripts can be easy to change.

Files can be written, so that record-oriented interviews and data-collection can be accomplished.

The interviewer can back up to any previous question and resume forward motion without re-entering the responses. This is true no matter how complex the questionnaire path may be.

2.4 The Authoring Language

A special authoring language and processor has been written to greatly simplify the process of creating interviews. The authoring script conducts an interview with the questionnaire designer to ask questions about the interview to be designed. The SPL code is then automatically generated to conduct the target interview. This simplifies the construction of interviews and helps to insure correctness of the resultant survey instrument.

Most of the required question types are available through the authoring language. Examples include, numeric responses, short text responses (such as name), long text responses (up to 20K of text with word-processing type editing available), and the following multiple choice types: 1) fixed list; 2) fixed list with an "other" alternative; 3) list taken from database values; 4) list that can be added to by the user; 5) list with more than one response permitted. The multiple choice lists can be stored either on the database or on a simple ASCII file. The authoring language covers testing with conditional branching and loops. Loops are required for questionnaire subsections (e.g., for each child, answer the following —). Hard-copy questionnaires and most interview systems require that limits be set for such loops (e.g., a maximum of seven children). HyperQuest imposes no such limits on the user, and wastes no disk space for questions not answered or for fields that are blank when the actual number of entries is less than the maximum number allowed.

3. THE HYPER-RELATIONAL DATABASE

This section describes the Hyper-Relational database concept and how it is different from conventional database models. Examples shown are not related to survey design so that no knowledge of survey techniques is required to understand this discussion.

3.1 The Hyper-Relational Concept

HyperQuest uses a new database model invented for this system: the hyper-relational database model. This model has the following advantages over conventional Relational databases:

Relations are defined at the data element (field) level. This eliminates data normalization problems inherent with maintaining tables of data.

Fields and relations can be added or deleted without requiring database reorganization.

Complex operations, such as joins, can be made with an object-oriented query system that does not require the user to know a query language such as SQL.

All entity-relationships are supported automatically. These include one-to-one, one-to-many, many-to-one, many-to-many and repeating subgroups.

Rollback, and re-indexing of data are simplified since only a few physical files are involved in the database.

The same database can contain a diversity of applications. This makes it easy to relate different applications or sub-problems without complex distributed systems.

No predefined schema are required — only a simple dictionary that defines the relationships between the data elements.

Missing data problems disappear, since missing data are not present in the database.

Queries are extremely rapid.

3.2 The Hyper-Relational Model

3.2.1 The Relation

Instead of the logical records common to most database systems, the Hyper-Relational model has the concept of the relation. A relation is used to store one occurrence of a data element and to relate that occurrence to another data element. A relation consists of an attribute and two arguments, thus forming a triple. Consider the following examples of relations:

LANGUAGE(JOHN SMITH, ENGLISH)

This relation asserts the LANGUAGE for JOHN SMITH is ENGLISH.

AGE (WALTER, 42)

The AGE of WALTER is 42.

MANAGES(AL,JOHN)

AL MANAGES JOHN.

This method of representing information has been used by Artificial Intelligence systems and the triples are called OAV triplets (Object, Attribute, Value), but that concept is considerably extended here through the use of symbols and objects.

3.2.2 Symbols

The arguments of a relation can be numeric values, text strings or symbols. A symbol is an argument which is a word or a name. Symbols are indexed and provide a handle for navigating the data base. Because of this indexing method, it is easy to find all persons managed by AL, all persons who manage JOHN, the AGE of all persons, all relations involving WALTER, etc.

3.2.3 Objects

An object is a collection of symbols that represent one type of real-world entity. In the following relations, for example,

LANGUAGE(JOHN SMITH, COBOL)
LANGUAGE(JOHN SMITH, ADA)
LANGUAGE(JOE BROWN, COBOL)
LANGUAGE(JOE BROWN, FORTRAN)

COBOL, FORTRAN and Ada are all programming languages. John Smith and Joe Brown are both names of employees. The set of programming languages can be thought of as an object and given a name, say <PROGRAMMING-LANGUAGE>.

Similarly, the set of employees can be thought of as an object and given a name, say <EMPLOYEE>.

An object contains one type of item. The <PROGRAMMING-LANGUAGE> object can be thought of as a box that can contain a set of cards. On each card is the name of one programming language. Cards can be added or removed from the box.

3.2.4 Relational Definitions

A data dictionary defines the relations in the database in terms of the objects they contain. For example, a dictionary entry for the LANGUAGE relation used above, would be:

LANGUAGE(<EMPLOYEE>,<PROGRAMMING-LANGUAGE>)

This line specifies that the <EMPLOYEE> object and the <PROGRAMMING-LANGUAGE> object are related through the LANGUAGE relation.

3.2.5 Database Navigation and Queries

Consider the following two relational definitions:

LANGUAGE(<EMPLOYEE>,<PROGRAMMING-LANGUAGE>)
PROJECT-LANGUAGE(<PROJECT>,<PROGRAMMING-LANGUAGE>)

The LANGUAGE relation defines which languages the employees know and the PROJECT-LANGUAGE relation defines which languages are used on projects. To find out which employees know the languages required by some projects, the following steps would be executed:

Create an <EMPLOYEE> object
from the <PROGRAMMING-LANGUAGE> object
from the <PROJECT> object
from a selection of projects.

The user can specify this query easily by selecting choices in a menu driven system. Execution is backwards:

the user selects projects of interest;
the programming languages used by the selected projects
are retrieved (PROJECT-LANGUAGE relation);
the employees knowing the resultant programming
are retrieved (LANGUAGE relation).
The result is a list of employees with the
desired attributes.

In general, objects can be used to make new objects (make list of employees who know COBOL and Ada), restrict objects (restrict the list of employees to those who have worked on certain projects), add to objects (add employees to the list of employees who have master's degrees), remove objects (remove employees from the list of employees if they worked on project x). These operations are equivalent to logical AND, OR and NAND operations, without confusing the user with such terminology.

3.2.6 Creating Relations

The interview process, either through the authoring language or SPL directly creates the relations that can be queried. For example, a marital status question can produce the relation:

MARITAL_STATUS(<respondent>,<status>)

Examples might include:

MARITAL_STATUS(345,SINGLE)
MARITAL_STATUS(345,DIVORCED)

Questions that involve multiple responses result in multiple relations. For example, a question asking for favorite flavors of ice cream can result in:

FLAVOR(345,VANILLA)
FLAVOR(345,CHERRY)
FLAVOR(345,PEACH)
FLAVOR(89,CHERRY)

Respondent 345 likes vanilla, cherry and peach while respondent 89 likes only cherry. A dictionary entry for this relation could be:

FLAVOR(<respondent>,<preferred_flavors>).

Queries such as, "which respondents like cherry ice cream?" or "what flavors does respondent 345 like?" can be made.

4. THE REPORTING MODULE

In addition to a query capability, the ability to obtain statistical information and reports is an essential component of a good survey system.

Statistical information, such as the mean, standard deviation, minimum and maximum can be obtained for numeric data using the query processor. An ability to analyze categorical data in this way is planned.

A report writer is a part of the query system. The user "paints" a picture of the report desired on the screen. Fields can be selected, headings and footings produced. Repeated groups can be defined and printed. Reports can be saved and retrieved as needed. It is possible to produce fairly complex reports very quickly and easily.

Sometimes, sequential records need to be produced for entry into another program. SPL can be used to write any data to sequential files in any format desired.

5. CONCLUSION

The merger of an interview engine designed expressly for survey work and a hyper-relational database with query and reporting functions represents a significant advance in computer-assisted interviewing technology. The goal has been to provide a powerful tool for the interview designer and an extremely simple tool for the interviewer. HyperQuest is not a static system. We are adding new features as required by actual applications, and suggestions from current and potential users are always welcome.

Matthew Jaro is Director of Research and Development
System Automation Corporation
8555 16th Street
Silver Spring, Maryland 20910
U.S.A.

(301) 565-9400

THE IMPACT OF HIGH TECHNOLOGY ON DATA COLLECTION

W. L. NICHOLLS II,

US Bureau of the Census

THE IMPACT OF HIGH TECHNOLOGY ON DATA COLLECTION

William L. Nicholls II
Field Division, U.S. Census Bureau[1]

This paper considers the impact of high technology on survey data collection. Although the paper examines a broad sweep of time, from 1970 to 2000, its scope is limited in two ways.

First, this paper is primarily concerned with the impact of high technology on governmental data collection. Private sector use of the same technologies is mentioned largely as a step toward their adoption by government data collection agencies.

Second, the paper reports primarily on developments in the United States and especially on those at the U.S. Census Bureau. This focus is not meant to minimize similar activities elsewhere but merely reflects the author's greater familiarity with activities nearer home.

The paper divides the impact of high technology on survey data collection into two phases.

Phase 1 consists of technological innovations introduced into survey data collection before 1980. The focus is on computer-assisted telephone interviewing (CATI), first used in the private sector in the 1970's. Brief mention is also made of two associated methodologies, random digit dialing (RDD) and centralized telephone interviewing.

Phase 2 technologies consist of innovations with their first major data collection applications after 1980. They include by-products of the PC revolution, such as computer-assisted personal interviewing. They also include such new developments as speech technology and portable handwritten character recognition devices which may find major uses in survey data collection by the year 2000.

An assessment of the emerging consequences of the Phase 1 technologies on survey data collection may better prepare us to meet the challenges of Phase 2 technologies.

1. Phase 1 Technologies

Computer-assisted telephone interviewing (CATI) constitutes the first major example of the impact of high technology on survey data collection. It applied interactive computing to data collection tasks performed since the inception of survey interviewing by paper-and-pencil methods. Not only does CATI combine data entry with data collection, it brings the power of interactive computing to the previously clerical tasks of field record keeping, call and callback scheduling, phrasing of questions to fit the situation, contingent probing, controlled branching between items, and the detection and clarification of inconsistent responses during the interview.

The consequences of CATI for survey data collection could be assessed in several ways already well covered in the literature. Many papers have described new data collection capabilities (or options) made possible by CATI [11, 13, 17, 29, 41, 54]. Others have examined CATI's effects on survey data quality, survey estimates, and survey costs [7, 17, 29, 32]. A few papers have considered CATI's impact on the administration of surveys and the internal structure of survey organizations [2, 4,

---

[1]/The views expressed in this paper are the author's and do not necessarily represent official views or policies of the U.S. Census Bureau.

9, 24]. This paper takes a somewhat different perspective. It assesses the impact of CATI through the status and pace of CATI implementation in government data collection agencies and by the consequences of, and impediments to, the use of CATI by those agencies.

1.1 Current and Projected Use of CATI by U.S. Government Agencies

How broadly is CATI used by U.S. governmental data collection agencies? The answer is suggested by Table 1 which reports the number of CATI interviewing stations and CATI sites (or installations) at four Federal agencies. Smaller installations at other U.S. agencies may have been overlooked. In this and in the following tables, the count of CATI stations includes supervisory and managerial stations as well as those used directly for interviewing. In some agencies the same stations also are employed for data entry from paper forms and other purposes.

The four U.S. agencies represented in Table 1 currently have almost 500 CATI stations in operation. If current plans at these same agencies are fully realized, then, as shown in Table 2, their combined total of CATI stations will approximately triple by 1994.

TABLE 1

CATI SITES AND CATI STATIONS IN U.S. FEDERAL AGENCIES, 1988

Agency	CATI Sites	CATI Stations	Stations Per Site
Census Bureau	2	104	52.0
Field Division	1	32	32.0
Data Preparation Division	1	72	72.0
Bureau of Labor Statistics	14	62	4.4
Consumer Price Index Surveys	1	8	8.0
Special OEUS Surveys	1	18	18.0
Current Employment Statistics Program	12	36	3.0
National Agricultural Statistics Service	14	193	13.8
National Centers for Disease Control	21	140	6.7
Total	51	499	9.8

A closer look at their plans reveals major variations across agencies in the use of CATI and its pace of implementation.

TABLE 2

CURRENT AND PROJECTED CATI SITES AND STATIONS
IN U.S. FEDERAL AGENCIES: 1988 - 1994

Agency	CATI (Sites) and Stations							
	Current		--------- Projections				----------	
	1988		1990		1992		1994	
Census Bureau	(2)*	98	(2)	138	(3)	210	(4)	282
Field Division	(1)	32	(1)	72	(2)	144	(3)	216
Data Prep. Division	(1)	72	(1)	72	(1)	72	(1)	72
Bureau of Labor Statistics	(14)	62	(14)	62	(32)	139	(53)	272
CPI Surveys	(1)	8	(1)	8	(1)	16	(1)	50
Special OEUS Surveys	(1)	18	(1)	18	(1)	18	(1)	18
CES Program	(12)	36	(12)	36	(30)	105	(51)	204
National Agricultural Statistics Service	(14)	193	(30)	375	(42)	747	(42)	747
National Centers for Disease Control	(21)	140	(21)	140	(26)	175	(26)	175
Total	(51)	499	(67)	721	(103)	1277	(125)	1482

*CATI sites in parentheses.

The Census Bureau currently has two CATI sites. The first is the Field Division's Hagerstown Telephone Center, which is used primarily for surveys of household residents and for small surveys of industries. This site will be upgraded from 32 to 72 CATI stations by 1990, but plans to add two additional CATI sites of comparable size in other parts of the nation by 1994 depend on expanded use of CATI in the Current Population Survey which are not yet approved.

The Census Bureau also operates a CATI facility through its Data Preparation Division in Jeffersonville, Indiana. This facility collects data from establishments for the Retail and Wholesale Trade Surveys of the Bureau's Business Division. Its 72 networked microcomputer work stations are used for CATI interviews, data capture from mailed questionnaires, and failed edit follow-up on the same surveys. There are no immediate plans to increase the size of this site, but several additional years will be required to move the remaining Business Division Surveys to CATI.

The Bureau of Labor Statistics (BLS) has three data collection programs using CATI. The first is a small headquarters test site developing CATI methods for the Consumer Price Index (CPI) Surveys. By 1994, it will be augmented by a national, 40- to 60-station CPI production interviewing facility. The second BLS CATI program uses an 18-station headquarters facility for special surveys of the BLS Office of

Employment and Unemployment Statistics. The third and largest test use of CATI by BLS is in the Current Employment Statistics (CES) Program. In this monthly establishment survey, CATI is used for interviewing, nonresponse follow-up, and failed edit reconciliation. If initial testing proves successful, BLS will expand CES CATI from 12 to 51 State offices by 1994.

The National Agricultural Statistics Service (NASS) currently operates the U.S. government's largest CATI network for its surveys of farms and farm operators. A procurement now in progress will upgrade all 14 NASS State offices already equipped with CATI and install CATI in all but two (Alaska and Hawaii) of its remaining State offices by 1991. Many of the NASS CATI stations also will be used for data entry of paper questionnaires, for failed edit follow-up, and for office automation functions.

The National Centers for Disease Control operates a network of 21 CATI-equipped State offices which conducts interviews for the Behavioral Risk Factors Surveillance Survey and related RDD household surveys [37]. Only modest growth is anticipated in the CDC CATI network over the next few years since data collection in many remaining States is contracted to local survey organizations.

Three general points can be made from these agency plans:

First, those U.S. government agencies now expanding their CATI facilities or surveys expect to complete those plans between 1990 and 1994. Not all these plans may be realized in full or on schedule, but the end of the basic implementation phase of Federal CATI data collection is in sight.

Second, the predominant use of in-house CATI facilities by U.S. government agencies will <u>not</u> be in support of large staffs of telephone interviewers centralized in national or regional offices. Rather, as shown in Table 3, CATI typically will be used to support smaller State-based staffs who often utilize the same equipment for data entry of mailed returns, failed edit reconciliation, and for related activities.

Third, even in those agencies exhibiting the most rapid increase in CATI facilities, CATI will not necessarily become the predominant mode of data collection. At NASS and in the BLS Current Employment Statistics Program, mailed questionnaires remain extremely important components of their mixed mode data collection methodologies. At the Census Bureau, the 280-300 CATI stations planned for 1994 will still be dwarfed by a field staff of 2,500-3,000 field representatives conducting personal interviews. For the Census Bureau, equipping this field staff with portable microcomputers for computer-assisted personal interviewing (CAPI) will be a much larger task.

1.2 Factors Affecting the Pace of CATI Implementation

With the end of basic CATI implementation now in sight for U.S. government agencies, we may reflect briefly on the factors affecting the pace of implementation. The first private sector CATI survey was conducted by Chilton Research in 1971; but even those governmental agencies moving most quickly will not fully implement CATI in their major continuing surveys before 1991. That will be 20 years, or a full generation, after CATI was invented. For the Census Bureau's major household surveys, such as the Current Population Survey and the National Crime Survey, the earliest conceivable date for full CATI implementation is 1994, but slippage, say to 1996, is possible. That would represent a quarter century after the first CATI survey in the private sector. Why has it taken so long?

Many different factors undoubtedly have contributed to this 20-to 25-year implementation period. In the early 1970's, the managers of large Federal household surveys, according to Dillman and Tarnai [9], tended to dismiss telephone interviewing

TABLE 3

PROJECTED CATI SITES, CATI STATIONS, AND STATIONS PER SITE IN U.S. FEDERAL AGENCIES, 1994

	Sites	Stations	Stations / Site
National & Regional Facilities	6	356	59.3
Census Bureau	4	288	70.5
BLS CPI & OEUS	2	68	34.0
State Sites	119	1,126	9.5
NASS	42	747	17.8
BLS - CES	51	204	4.0
CDC	26	175	6.7
U.S. Total	125	1,482	11.9

as an inferior method of data collection. Where a readiness for change was present,the technology often was lacking. CATI software did not have the features needed for many governmental surveys until they were developed by university organizations (with government support) in the late 1970's [40]. After 1980, when government agencies began active internal development of CATI, they often initiated major research programs to assess CATI's effects on costs, data quality, and estimates. At the U.S. Census Bureau and Statistics Canada, these programs are only now producing final reports [7, 57]. The research has been especially thorough where a change of collection methods could impact policy-sensitive time series. Like private survey organizations, government agencies have moved more quickly to introduce CATI in new or infrequently conducted surveys. The long lead times and frequent delays in government planning, budgeting, and procurement have undoubtedly further slowed the pace of government CATI implementation.

1.3 CATI, RDD, and Centralization

CATI's extended incubation period in government also may be partly explained by its initial association with two related methodologies, random digit dialing (RDD) and centralized telephone interviewing [50, 59]. Together, RDD, centralized telephone interviewing, and CATI are sometimes described as "modern telephone methods." Their joint evolution has been described by Groves and Kahn [14] as one of the major developments in the history of survey methods, ranking with area probability sampling and the use of computers for survey analysis. By 1980, most U.S. university survey research centers and market research organizations had established central telephone facilities conducting RDD and other surveys [4, 32, 48]. Acquisition of CATI frequently followed [49]. Government agencies were the exception.

"Modern telephone methods" did not transfer readily to government data collection as a package. This is most apparent for random digit dialing, whose potential to reduce survey costs attracted major interest among government statisticians. The U.S. National Center for Health Statistics [53], the U.S. Census Bureau [3, 26, 45], and Statistics Canada [10] all began their investigations of modern telephone methods with years of careful testing of random digit dialing [27]. As Drew, Choudhry, and Hunter [10] have observed, however, RDD sampling methods are used in few government surveys conducted in the U.S. or elsewhere in the world. The omission of nontelephone households (about 7 percent of the U.S. total) and the typically higher refusal rates of cold contact telephone interviews have presented major barriers to the use of RDD in many or most government survey applications.

Random digit dialing remains a valuable sampling method for populations with high telephone subscribership and for surveys which can tolerate its coverage and nonresponse problems. For some governmental statistical agencies, however, the early emphasis on RDD proved a diversion from what now appear to be more fruitful uses of CATI. Only when RDD was ruled out as a sampling method for most U.S. government household surveys, which at the Census Bureau occurred around 1986, could plans to implement CATI in single-frame, mixed mode designs proceed. The somewhat faster adoption of CATI by establishment surveys may be partly attributable to their traditional reliance on list frame samples. A change to RDD was not an issue. Nor was RDD sampling apparently an issue in implementing CATI in the Centers for Disease Control's Behavioral Risk Factors Surveillance Survey [37]. RDD was chosen for this survey from the start.

The second major element of modern telephone methods which has not translated easily to government data collection is centralized telephone interviewing. In U.S. university and commercial market research, the shift from "dispersed" local interviewers making calls from their own homes to "centralized" telephone interviewers calling from national or regional offices was largely completed by the late 1970's. Only government household surveys continued major use of dispersed telephone interviewing from interviewers' homes into the 1980's.

Centralized telephone interviewing and CATI are linked both historically and as mutually reinforcing methodologies. Chilton Research pioneered centralized telephone interviewing in 1966 [22] and then introduced CATI as its automated form five years later [29]. Computer-assistance is easier to arrange for centralized interviewers who share the same hardware, programs, sample, and technical staff [15, 16]. At the same time, CATI encourages centralized interviewing to gain these efficiencies and to benefit from such large-staff CATI features as automatic call scheduling, online supervision, and field report generation [9, 16, 30]. Centralization contributes to standardized field procedures and interviewing quality control through easier recruitment, training, and supervision of interviewers, while CATI contributes to these same goals through tailored question wordings, computer controlled branching, and online editing. Supervisory audio-visual monitoring of interviewer performance, currently feasible only with centralized CATI interviewers, provides feedback ensuring that CATI quality enhancement features are used appropriately and that interviewers deviating from performance standards are identified and retrained when necessary.

Centralized telephone interviewing also has potential disadvantages. If the role of interviewer is too specialized, interviewer morale may be impaired [26]; and centralized interviewers do not have the local accents, knowledge, and cultural understanding which local interviewers may share with respondents [24]. When a mixed mode design requires centralization of only part of a survey's data collection, unusually effective management, coordination, and communications are required [2]. And if centralization means transferring existing work from one staff to another, major staff or organizational resistance may be encountered [9]. This is especially true in continuing government surveys which frequently employ interviewers,

supervisors, and managers with many years of service.

The most difficult problems of implementing CATI in government agencies appear to derive from the organizational issues CATI typically raises about centralized vs. decentralized interviewing. At Statistics Sweden, Lyberg [24] reports that upper management settled this question early by making a commitment to dispersed telephone interviewing from interviewers' homes, while limiting centralized CATI to small, specialized surveys and nonresponse follow-up. Traditional interagency or Federal-State (Province) agreements, formal or informal commitments to employees, and priority policies of the current administration may prove as important to these decisions as the goals of maximizing error control and operational efficiency.

The centralization issue takes a somewhat different form for establishment surveys. Establishment surveys conducted by government agencies generally began as mail shuttles using mailout/mailback methods. Telephone interviews typically were first introduced to prompt mail returns, accept late replies by phone, and to reconcile edit failures. For efficiency, the same office staff may handle both mail returns and telephone interviews. The key organizational issues are whether mail and/or telephone operations should be located in national, regional, or State (Province) offices.

The introduction of CATI strengthens the arguments for centralization. The Census Bureau's Business Division took this path in preparing to automate its Retail and Wholesale Trade Surveys. Telephone interviewing was withdrawn from the Bureau's 12 Regional Offices and centralized at the Data Preparation Division's Jeffersonville office where mailing, data capture, and initial processing were already located. An integrated system combining frame updating, mailing, CATI, data capture, editing, and data base operations could then be introduced more easily and a centralized staff more easily trained and supervised in its use. Since this change moved work from one division to another, it was not accomplished without occasional strain and close administrative scrutiny, but the decision undoubtedly facilitated automation of these surveys.

When surveys are too large to be accommodated by a single national CATI facility, regional offices are the nearest alternative. The National Agricultural Statistics Service considered such a structure for its surveys of farms and farm operators. In one plan, each of the 41 State offices would be responsible for its own face-to-face interviews, but only a few State offices would have administratively attached CATI facilities to conduct telephone interviews in their own and nearby States [2]. A major test of this plan was undertaken to determine whether NASS offices could conduct CATI interviews effectively outside their own States. Neither the location of the CATI facility nor its level of prior CATI experience were found to affect data quality or key estimates [2]. The research foreclosed neither regional nor State locations for NASS CATI facilities. All 41 State offices will be equipped for CATI although some surveys may be conducted from a smaller set.

Centralization of telephone interviewing may be more important to household surveys than to establishment surveys. Werking, Tupek, and Clayton [60] have suggested that response errors in household surveys derive primarily from such sources as faulty recall, telescoping, and memory errors. These are the kinds of problems which may be reduced by close adherence to good interviewing practice encouraged by extensive interviewer training, supervision, and monitoring. By contrast, the principal sources of response error in establishment surveys apparently stem from the establishments' record keeping systems. Reduction of these errors may be less dependent on standardized interviewer performance than on periodic record checks to clarify the exact type of information needed from responding establishments. This argument supports what now appears to be the predominant usage of CATI by U.S. government agencies, that is for establishment surveys conducted from relatively small, decentralized State offices.

What are the primary benefits of CATI for such establishment surveys? Tortora [54] and Werking et al., [60, 61] have cited such familiar benefits as: controlled question branching; online range, consistency, and longitudinal editing; data capture concurrent with data collection; reductions in the need for fail-edit callbacks; and contributions to survey management. Neither has claimed reduced survey costs with CATI. Werking and his BLS colleagues also have cited gains in response rates, reductions in panel attrition, and improved timeliness of results as major gains in substituting CATI for mail collection methods in the Current Employment Survey Program. These later benefits, however, may derive directly from the substitution of telephone for mail data collection methods rather than from CATI technology *per se*.

1.4 Implementing CATI in Household Panel Surveys

The implementation of CATI in government household surveys frequently poses even more design and organizational problems than its implementation in establishment surveys. This is illustrated by the Census Bureau's plans to phase CATI into three national household panel surveys: the Current Population Survey (CPS); the National Crime Survey (NCS); and the American Housing Survey (AHS). Each of these household panel surveys has a mixed mode design. The first visit to each sample address is by personal visit to identify ineligible housing units and to encourage household participation. The fifth CPS and NCS visits also are in person to re-establish personal contact with the household part way through the sequence of interviews. Other interviews are by telephone when possible and acceptable to the respondent and by personal visit otherwise.

The same local interviewers currently conduct both the personal and telephone visit interviews, placing the telephone interview calls from their own homes. Since the CPS and NCS are conducted every month, they have permanent staffs of interviewers. The National AHS is conducted only every other year; and its interviewing staff must be rebuilt for each wave. Local interviewers frequently work on more than one Census Bureau survey each month and this permanent staff generally provides the core of interviewers for less frequently conducted surveys. The interviewers are supervised by 12 Regional Offices.

When CATI is introduced into these surveys, no change is made in the initial visits to each sample address. These remain personal visit interviews since comparable response and panel retention rates have not been attainable with cold contact telephone interviews [26]. CATI replaces dispersed telephone interviews from the local interviewers' homes in the second and later visits of the panel design. This field design has several potential benefits: (1) reduced field costs; (2) reduced interviewer recruitment problems in tight labor markets; and (3) possibly improved survey estimates as described in the next section [21, 34, 57]. Nevertheless, the transition poses a number of design and organizational problems which require time and effort to resolve.

The first is developing appropriate methods for rapid but controlled transfer of individual case records between personal visit and CATI interview modes. When the first-visit personal interview is complete, the household enumeration data and field records must be data entered into computer files for second and later interviews by CATI. Case records also move from CATI through the Regional Offices to the local interviewers for CPS and NCS fifth visits and for personal follow-up of households unreachable by CATI. Procedures have been developed to accomplish these transfers within the required time frames, but the process of moving cases back and forth between dispersed and centralized interviewing staffs (and between paper records and computer files) is a complex one vulnerable to error. Conversion to an integrated CATI-CAPI system, using computer-assisted personal interviewing in the field, should

facilitate more rapid and reliable transfers.

The second transition problem in introducing CATI into household panel surveys is the temporarily reduced efficiency of the sample designs. Both the CPS and NCS employ cluster samples chosen mainly to minimize costs for interviewing assignments containing both personal visit and dispersed telephone interviews. When the dispersed telephone interviews are removed to CATI, the remaining personal visit cases may no longer constitute acceptable or efficient field assignments. In single-interviewer primary sampling units (PSUs), the telephone interviews simply cannot be transferred to CATI without reducing the interviewers' workloads (and compensation) below acceptable levels. In multiple-interviewer PSUs, the personal visit interviews may be redistributed after the telephone visit cases are removed, but as fragments of previous clusters, they may be poorly distributed for efficient travel by a smaller number of interviewers. The obvious solution is a revised sample design which provides efficient field assignments after most telephone interviews are transferred to CATI; but the CPS and NCS samples are based on the decennial census and revised only once a decade. A newly designed sample cannot be fully in place before 1995.

The third problem in moving dispersed telephone interviews to CATI is the need to reduce the field staff while increasing the CATI staff. For the CPS, the Census Bureau's largest current survey, the transition will be based initially on field interviewer attrition [57]. Cases will be sent to CATI only as CPS field interviewers separate from the Census Bureau for other reasons. Both the pace and magnitude of CATI implementation will therefore be initially constrained by the rate of interviewer attrition in multiple-interviewer PSUs.

The fourth and final transition problem is finding a sufficient volume of work for the CATI interviewers. The CPS conducts its interviews in the third week of each month and the NCS in the first week with some carryover into the second. Centralized CATI interviewing is restricted to even fewer days per month to permit field follow-up of cases unreachable by telephone. These two surveys will provide the CATI staff with relatively few days of employment per month. The problem is most severe for the CPS. Of the three CATI facilities the Census Bureau is planning for 1994, two would be required almost exclusively for the CPS. Additional CATI surveys providing more work hours, a change in the CPS field work design, or a cap on the number of CPS interviews sent to CATI may be necessary for an economically viable mixed mode use of CATI in the CPS.

Of the four transition problems just listed, only the first derives from the CATI technology. The transfer of cases between dispersed local interviewers and centralized CATI interviewers is complicated by the move between paper-and-pencil records and computer files. The problems of field sampling efficiency, field staff phase-down, and insufficient work at the CATI facilities arise from the centralization of previously dispersed interviews. They would be the same whether the central facility used CATI or paper-and-pencil methods. The primary problems of implementing CATI in government household surveys derive not from its technological but from its organizational requirements.

1.5 Some Recent Results

Although the introduction of CATI (and centralized telephone interviewing) into continuing government household surveys presents a number of design and organizational problems, there is increasing evidence that these methods may yield important, if occasionally disquieting, benefits. The examples presented here are from the Census Bureau's evaluative studies of CATI in mixed mode designs. They differ in two important respects from the evaluations of CATI conducted by Catlin, Ingram, and Hunter at Statistics Canada and reported in the following paper of this symposium [7].

First, Catlin, Ingram, and Hunter employed a research design comparing centralized CATI with centralized paper-and-pencil telephone interviewing. Observed differences generally can be attributed to the use of CATI rather than paper-and-pencil methods. The studies conducted by the Census Bureau compare mixed mode data collection using centralized CATI with mixed mode data collection using dispersed paper-and-pencil telephone interviewing. Thus, they cannot readily distinguish the effects of CATI from the effects of centralized telephone interviewing.

Second, Catlin, Ingram, and Hunter focus their analysis primarily on cross-sectional measures of data quality, such as unit and item response rates, item consistency within the same interview, and the codeability of occupation and industry entries. The Census Bureau results presented here focus on survey estimates in panel surveys.

Schwanz, Montfort, and Cannon [42] examined the use of CATI in the National component of the American Housing Survey, which is conducted biennially in odd numbered years. In 1987, two of the six AHS sample panels were assigned to CATI to assess its utility in panel interviews with selected types of previously interviewed housing units. Although 39.5 percent of the CATI-assigned housing units could not be reached by telephone and had to be recycled to the field, the investigators concluded that first attempts by CATI yielded savings in both time and money. They recommended similar use of CATI in all six AHS National panels in 1989.

This study also examined the value of computer-assisted reconciliation interviews in estimating change in housing characteristics. When the interview was conducted by CATI, the 1987 replies to nine key items were compared with the same household's responses to these items in 1985. The 1985 values were not accessible to the CATI interviewer until the last response to the 1987 interview had been entered. When difference between the 1985 and 1987 replies exceeded programmed tolerance limits, a reconciliation question was displayed asking the reasons for the discrepancy. This technique is familiar in establishment surveys, where it is called "longitudinal editing" or the "use of historical data" [35]. It is less common in household surveys where the same procedure is called "dependent" or "reconciliation" interviewing.

More than half the CATI AHS interviews (54.8 percent) failed one or more of the longitudinal edits and triggered at least one reconciliation item. The reconciliation responses suggest that about half the identified differences represented actual changes in housing characteristics and about half were response errors. If changes in housing characteristics were estimated from conventional cross-tabulations of 1985 and 1987 responses to the same items for the same housing units, about half the estimated change would be spurious.

Table 4 provides examples. In the CATI interviews, 19.2 percent of the households reported a different type of home heating equipment in 1987 than in 1985. When confronted with the difference, however, the reported change was reduced to 5.7 percent while the respondents attributed the remaining 13.5 percent to errors in the 1985 or 1987 reports. Similarly, while 12.0 percent of the households reported a different type of basement in 1987 than in 1985, actual change appears to have been only 1.6 percent while the remaining 10.4 percent were reporting errors. Curiously, errors in the respondent's own 1987 replies, provided only minutes before, were reported as often as errors in the two-year old responses from 1985. Such longitudinal editing may be essential to estimate change over time in the characteristics not only of housing units but also of household members.

Although longitudinal editing is facilitated both by CATI and by centralized telephone interviewing, it undoubtedly could be carried out with dispersed computer-assisted personal interviewing (CAPI) in the field. Pafford and Coulter [35] have urged that historical data not be used in paper-and-pencil interviews where prior interviewer knowledge of previous responses may invite both response biases and

TABLE 4

REPORTED SOURCES OF DIFFERENCE BETWEEN SELECTED 1985 AND 1987 HOUSING CHARACTERISTICS AMERICAN HOUSING SURVEY CATI SAMPLE

Housing Item	Total 1985-1987 Difference	Reconciliation Actual Change	Reconciliation Reported Error
Heating equipment	19.2%	5.7%	13.5%
Number of bedrooms	12.3	7.6	4.6
Kind of basement	12.0	1.6	10.4
Home value	10.2	7.7	2.5
Fuel used	9.9	5.2	4.6
Number of bathrooms	8.3	1.9	6.4

interviewer variability in their use.

Hubble and Wilder [21] analyzed the effects of CATI (and/or centralized telephone interviewing) on estimates from the National Crime Survey. This survey interviews a rotating panel of about 10,000 households per month who are interviewed at six month intervals. Each person, twelve years of age and over, living in the sample household is asked screen questions about personal crimes during the six months before the month of interview. The first respondent in the household also is asked screen questions about household crimes during the same recall period. If a respondent answers Yes to any of the screen questions, the interviewer completes a crime incident report for each incident. If no screen questions are answered with a Yes, the interview is quite brief.

The NCS telephone interviews are being phased into CATI beginning with multiple-interviewer PSUs in hard-to-enumerate areas. The design permits comparisons of CATI results with a control group that continues use of dispersed telephone interviewing. In 1987, when the figures in Table 5 were obtained, 200 to 400 households per month were assigned to the CATI test group. Not all test group interviews were by CATI. Personal interviews are required for the first and fifth semi-annual visit to each sample household. (The first visit data are employed only for bounding purposes, and estimates are based on data from visits two through seven.) Personal interviews also occur in other months in sample for households without telephones, those for whom a telephone interview is unacceptable or inappropriate, and for cases recycled to the field when unreachable by CATI.

A major objective of the NCS CATI research is to assess the effects of the mixed mode use of CATI on NCS crime rates. Since crimes are generally believed to be underreported, higher rates of reporting are typically assumed to be better. As shown in Table 5, the test methodology with CATI obtained significantly higher rates of both personal and household crimes. The personal crime rate was 29 percent higher for the test group than for the control group. For household crimes, the test group rate was 13 percent higher.

TABLE 5

ESTIMATED CRIME RATES PER 1,000 POPULATION
IN NCS CATI EXPERIMENT, 1987

Type of Crime and Interview Mode	Crime Rate Per 1,000 Persons		
	Test Group With CATI	Control Group	Ratio T/C
Personal Crime			
Total	113.4	87.9	1.29 **
Personal Visit	108.6	116.4	0.93
Telephone Visit	115.5	73.9	1.56 **
Household Crime			
Total	194.8	172.4	1.13 *
Personal Visit	185.8	194.8	0.95
Telephone Visit	198.9	161.1	1.23 *

*Statistically significant at the .10 level.
**Statistically significant at the .05 level.

Telephone interviews were the source of the overall difference. Crime rates from the test and control personal visit interviews did not differ significantly; but telephone interviews in the test group yielded personal crime rates more than 50 percent higher than in the control group. Although not shown in separate tables, the effect is not confined to less serious or to borderline crimes, but it does appear stronger for crimes unreported to the police.

This result is not necessarily produced by CATI. It may be a product of the opportunities centralized telephone interviewing affords for direct supervision, monitoring, and (when necessary) retraining of the interviewers. Unsupervised field interviewers calling from their own homes may not adhere as closely to the specified survey procedures. Or, both CATI and centralized telephone interviewing may contribute to the effect.

The causes of this effect are of more than academic interest. If computer-assistance improves crime reporting, then CAPI-equipped field interviewers calling from their own homes should do as well. If centralized telephone interviewing is the main cause, then centralized telephone interviewers using paper-and-pencil methods should do as well. The safest choice is to use both. The Census Bureau and the Bureau of Justice Statistics have agreed to transfer as many NCS telephone interviews to CATI by 1991 as the current NCS sample design allows. Future tests of CAPI in the NCS may tell us more about the relative importance of computer-assistance and centralized interviewing on this effect.

Bushery and Cahoon [6] have provided a third result of interest in their analysis of the effects of mixed mode CATI data collection on labor force estimates from the Current Population Survey. While this finding is not as dramatic as CATI's apparent effect on crime rates, it could prove more broadly relevant if supported by further research.

The pattern now known as the rotation group bias was first noticed in the Current Population Survey in 1955. Bailar [1] describes the phenomenon as a common one in panel surveys: the larger the number of times panel respondents are exposed to a survey, the less likely they are to report many of the things the survey was designed to measure. In the CPS, the general pattern has been much the same since 1970. Rates of employment, unemployment, and labor force participation decline from the first to the third month in sample. An additional survey question first asked in the fourth month produces a small upturn in these statistics in that month. The pattern is repeated, starting at a lower level, when the same households are returned for a second set of four interviews a year later.

Figure 1 compares the CPS civilian labor force rate across the first four months in sample for: (1) a test group using CATI for most of its telephone interviews in sample months two through four; and (2) a control group primarily using dispersed telephone interviewing during the same months. Both the test and control groups used personal visit interviews for the first month in sample and for nontelephone and other special households in months two through four. The control group illustrates the classic CPS rotation group pattern, while the test group has a significantly different pattern. When CATI replaces dispersed telephone interviewing, the characteristic CPS rotation group bias appears to be dampened.

We will not pause to consider other differences between the test and control groups in these preliminary results, such as the differences in level as well as by month in sample. But if further research continues to find that CATI (or centralized interviewing) dampens the rotation group bias, it may suggest that this well known phenomenon is at least partly a function of field interviewing practices rather than solely a consequence of respondent characteristics which change with repeated exposure to the same survey questions.

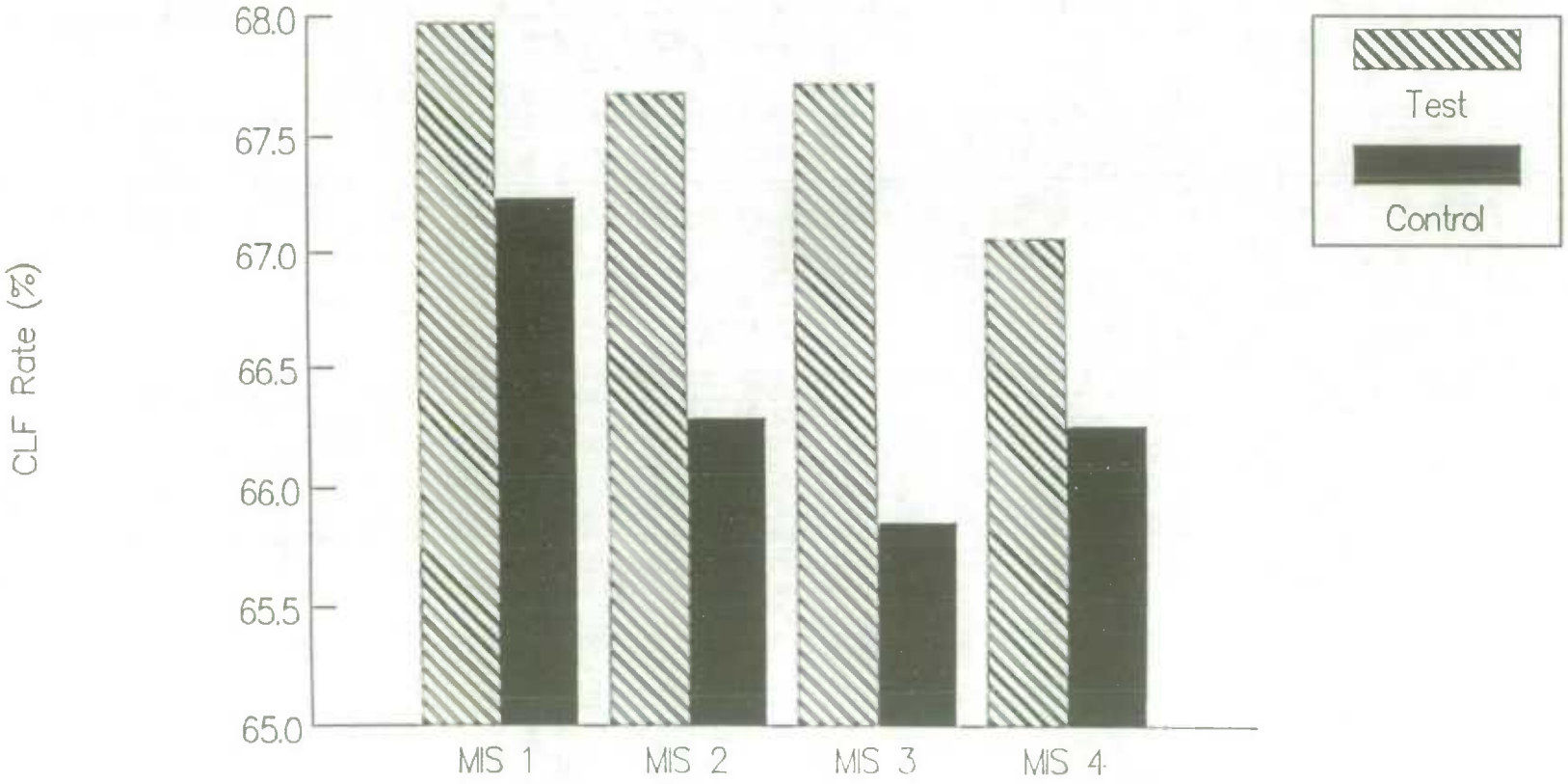
Again, we should be cautious about attributing this apparent finding to CATI. It may be a consequence of the close supervision of interviewer performance possible in centralized telephone interviewing. In either event, results from the AHS, NCS, and CPS may have opened new avenues of investigation deserving of serious attention. Computer-assisted telephone interviewing and centralized telephone interviewing may not merely be technological conveniences to facilitate routine survey data collection. They may prove to be major innovations which change our perspectives on the nature of survey data collection.

2. Phase 2 Technologies

The Phase 2 technologies are those whose first major use in survey data collection occurred after 1980 or which may occur before the year 2000. They include, but are not limited to, products of the PC revolution. Since little or no evidence yet exists about their impact on survey data collection, we can only describe the most promising technologies and speculate about their probable effects. This paper divides the Phase 2 technologies into four areas: (1) establishment survey technologies; (2) positional and communication technologies; (3) computer-assisted personal interviewing and related methods; and (4) voice technology.

FIGURE 1

Average CLF Participation Rates November 1986 to March 1988



2.1 Establishment Survey Technologies

While new (post-CATI) data collection technologies are applicable both to household and to establishment surveys, they are perhaps more easily introduced in the latter. Establishment surveys often require only brief numeric responses from the same respondents at frequent intervals. These expert respondents may welcome new data collection approaches utilizing technology familiar in business, especially when they minimize demands on their time.

The Bureau of Labor Statistics is successfully experimenting with voice simulation and touchtone data entry in its Current Employment Statistics Program [60, 61]. This system parallels those now used by many banks for phone-in account transactions. When respondents have prepared their monthly CES reports, they dial a number at a BLS local office where they are greeted by a voice simulation module and asked to touchtone enter their coded identification number. They are then asked a series of questions requiring numeric touchtone entries, confirmed with voice simulation as necessary. Since the procedure operates 24 hours a day without use of a telephone interviewer, its costs are minimal. Respondents who fail to call by the cutoff date are, of course, telephoned by a BLS interviewer. The next phase of this project will examine speech recognition modules to replace touchtone entry.

The Energy Information Agency (EIA) is exploring a different approach using computerized self-administered questionnaires (CSAQ) [23] in its Petroleum Supply Reporting System [52]. Visits to a small sample of respondents found that most were already using PC's available in their own or a nearby office. Diskettes were prepared for each company containing software to provide: (1) a computerized self-administered questionnaire to be answered each month; (2) menu driven procedures to assist the respondent in obtaining the needed information from other files or networked computers; and (3) programmed procedures to dial the EIA offices when the questionnaire is complete, enter the appropriate password, and transmit the data to the EIA computer.

Statistical agencies are not alone in devising new methods of data transmission for establishment surveys. At least one new approach apparently was invented by the respondents: using a facsimile (FAX) machine to transmit a copy of the paper questionnaire to the data collection agency. Large companies may find it convenient to transmit lengthy data sets on computer tapes.

As computer and communications technologies continue to evolve, other methods of collecting and transmitting data undoubtedly will be found. In the long run, the best data collection strategy for establishment surveys may prove to be a readiness to accept whatever combination of methods the respondents find most convenient. The CATI interview can be reserved as the backup for nonresponse, late returns, and inconsistent replies.

2.2 Positional and Field Communication Systems

Technological advances in the 1970's contributed most to telephone interviewing. Technological advances in the 1980's and 1990's seem likely to contribute most to personal visit interviewing. They may include both technologies capturing personal interview responses and technologies adding to the type of data collected.

In planning the 1990 Census, the Census Bureau reviewed a number of position-recognizing or navigational systems to ascertain whether low-cost, car-mountable, equipment was available which would provide reliable and repeatable geographic coordinates within 30 feet (or roughly 10 meters) of accuracy. Such equipment would have many uses in censuses and surveys, including: (1) map updating, such as adding a street to an online map by driving through it; (2) recording the precise coordinate position of rural structures to facilitate field callbacks; (3) geocoding precise

dwelling locations to permit grouping and tabulation of Census data to any geographic area definable by coordinate boundaries; and (4) use of similarly flexible geographic areas, unaffected by changes in roads, power lines, and political boundaries, for sampling purposes.

Systems approaching some of the desired capabilities were identified and closely examined. They included:

LORAN-C (for long range navigation system) uses low-frequency radio signals to determine one's position on the earth's surface [18]. The primary use of LORAN-C is for marine navigation, but car-portable units are available for terrestrial use. Unfortunately they did not prove sufficiently accurate and reliable to meet the Census Bureau's needs when last assessed in 1983. The LORAN-C system also omits Alaska, Hawaii, and the central portion of the United States.

NAVSTAR GPS (for global positioning system) is a satellite-based radionavigation system being developed for the U.S. Department of Defense. When complete, it will provide three-dimensional positioning anywhere on the earth [25]. Although the NAVSTAR "precise positioning service" will reportedly have 16-meter accuracy, its uses will be restricted to defense purposes. A "standard positioning service," available to a broader set of users, will have 25-30 meter accuracy. The NAVSTAR GPS system was scheduled to be operational in 1989; but it has been delayed by the long interruption in space shuttle operations.

The ETAK system (a Polynesian word rather than an acronym) is a car-mounted navigational system incorporating a digital magnetic compass, motion sensors, and a "smart" dead reckoning system [47]. It traces the car's position on a scrolling map displayed on a VDT mounted near the driver. Digitized maps of the local area, stored on magnetic tapes, are required, and cassettes must be changed as the driver moves from one area to another. Although such devices have been installed in emergency vehicles, they are not yet in common use by businesses, where the next stage of application would be anticipated.

When last examined by the Census Bureau, none of these systems were sufficiently accurate, low cost, and reliable for typical survey applications. However, since their accuracy, accessibility, and affordability will undoubtedly improve with time, they remain intriguing options for the future. For example, Magnavox representatives predicted in a presentation at the Census Bureau that by the year 2000 lightweight, portable NAVSTAR GPS receivers with accuracy to within 5 meters should be available for under $500 U.S. at local mountaineering supply stores. Car-mounted models with digital map displays may also be available from automobile electronic supply dealers by that time. Equipping interviewers with car telephones and/or paging beepers also may prove effective in the future, if their costs decline. The common survey practice of having interviewers drive their own cars will remain an impediment, however, to use of any car-mounted technologies.

2.3 Computer-Assisted Personal Interviewing

Computer-assisted personal interviewing (CAPI) is the most promising Phase 2 technology thus far introduced for household surveys. The concept of interviewing respondents in their own homes or offices using portable computers has been recognized for over a decade, but its practical application was limited until advanced handhelds and portable personal computers became commonly available.

Although small tests of CAPI had been conducted previously [5, 8], this technology first became a demonstrably viable method of national household data collection in 1987. In Europe, the Netherlands Central Bureau of Statistics began operational use of handheld computers in 1987 in the Netherlands Labor Force Survey [28, 58]. In the U.S., National Analysts began nationwide CAPI data collection for the USDA-sponsored 1987 Nationwide Food Consumption Survey [39]. Major tests of CAPI

also were conducted by the U.S. Census Bureau [31, 38], the U.K. Office of Population Censuses and Surveys, and the Research Triangle Institute [44].

In cooperation with the National Center for Health Statistics, the Census Bureau began continuing use of CAPI in the National Health Interview Survey (NHIS) in 1988. All 200 NHIS interviewers have been equipped with Gridlite Plus laptop computers and trained in their use. The first application is limited to a 15- to 20-minute NHIS supplement on AIDS. Additional NHIS supplements will be transferred from paper to CAPI beginning in 1990. Conversion of the core NHIS items to CAPI is anticipated in the future. The Census Bureau also is planning a major feasibility test of CAPI in the Current Population Survey in 1990 [55]. Plans have been drawn, although not yet approved and funded, to equip all CPS and NCS interviewers with CAPI in the mid-1990's [56].

The historical development of computer-assisted personal interviewing (CAPI) differs from the historical development of computer-assisted telephone interviewing (CATI) in at least two important respects.

First, CATI was initially developed in the private sector for market research surveys, and a decade elapsed before governmental agencies first conducted their own CATI surveys [32]. By contrast, government agencies are very much in the forefront of CAPI development, both as sponsors of CAPI tests by university and private sector organizations and in developing CAPI systems for their own use. The importance of personal visit interviewing to key Federal surveys and the relatively large investments required to initiate CAPI data collection may account for this.

Second, the implementation of CAPI in government surveys is proceeding at a much faster pace than occurred for CATI. CATI was already a relatively mature and well tested-data collection method before U.S. government agencies began serious consideration of its use. Even then, agencies such as the Census Bureau, Statistics Canada, and the then USDA Statistical Reporting Service (now NASS) conducted careful evaluations of its effects on data quality, survey estimates, and costs before moving toward production use [17, 32]. By contrast, implementation of CAPI in government surveys is proceeding rapidly without prior evaluative research. Feasibility tests have demonstrated that CAPI data collection is acceptable to almost all respondents and that most experienced field interviewers can be trained in its use [5, 28, 31, 38, 39, 44, 58], but thorough assessments of CAPI's effects on data quality and estimates have not yet appeared.

Key technological questions also remain currently unanswered for CAPI. First, "What are the best methods of transmitting assignments and questionnaires to CAPI interviewers and completed interview data back to the home office?" For the Nationwide Food Consumption Survey, National Analysts relied on the mails, UPS, and expedited courier services for the transmission of both assignments and completed interviews [39]. Completed interview data were returned on diskettes. National Analysts considered the use of telecommunications, with interviewers connecting their microcomputers to their home telephones at night, but rejected this approach as too risky, especially in non-Bell areas. For the first nationwide use of CAPI, mailing of diskettes was undoubtedly the wiser choice.

By contrast the Research Triangle Institute has used automated telecommunications successfully in two local tests of CAPI in the Total Exposure Assessment Method (TEAM) Survey conducted for the U.S. Environmental Protection Agency [44]. The Netherlands Labor Force Survey also is using telecommunications for transmission of CAPI data with reported success [58]. There are many advantages to rapid telecommunications between headquarters and the interviewers, especially for surveys operating on tight deadlines within brief interviewing periods, such as the U.S. Current Population Survey and the Canadian Labour Force Survey. The Census Bureau plans to develop an integrated CATI-CAPI system in which cases can be rapidly transmitted back and forth between CATI interviewers located in centralized sites and CAPI interviewers dispersed

across the nation. Whether this is feasible for CAPI interviewers stationed in every part of the country remains to be demonstrated. Transfer of data by telecommunications also poses additional problems (real and perceived) in the protection of confidential data. For first production use of CAPI in the NHIS AIDs supplement, the Census Bureau is following the more cautious approach of mailing diskettes.

A second unresolved question is: "How can CAPI software be made most convenient for the special requirements of household interviewing?" The software used in most CAPI applications to date was designed for computer-assisted telephone interviewing or for personal interviewing in office settings. Question formats, modes of entry, and interviewer movement commands acceptable in these more controlled settings may not be equally appropriate for use in the more demanding and distracting settings of respondents' homes or for doorstep interviews. Although general guidelines for computer-assisted questionnaires have been proposed [19, 20, 33], specific software features most helpful to CAPI interviewers have not yet been identified or widely shared.

A third unresolved question for CAPI, and perhaps the major one at present, is "With what type of hardware should CAPI interviewers be equipped?" While the Netherlands Central Bureau of Statistics developed its CAPI system for handheld micro-computers with small display screens, including the Epson PX-4, U.S. tests of CAPI generally have employed IBM compatible laptop computers, such as the Toshiba T-1100+ and the Gridlite Plus. These laptops have programmable function keys, full-sized screens and keyboards, and typically weigh at least 10 pounds. Current laptops with larger RAM, hard disks, and faster processing speeds, which may be necessary for questionnaires in two languages (French-English or English-Spanish), interviewing on different surveys on the same trip, and maintenance of call records, may weigh up to 16 pounds. This is an unacceptable weight for field interviewers. Their costs also are currently prohibitive for most survey applications.

Whether most interviewers can tolerate the burden of even a 10-pound laptop as a permanent part of their job is presently unclear. Reported experiences from recent tests of CAPI are inconsistent [31, 38, 44]. The issue may be resolvable only with experience, or by the availability of high performance portable microcomputers of lighter weight.

Off-the-shelf laptops also have other disadvantages for field interviewing. First, they are vulnerable to impact, spilled liquids, excessive dust or moisture, and extremes of temperature. Second, as general-purpose personal computers, interviewers (or members of their families) may be tempted to use them for other purposes than completion of interviewing assignments. This might compromise respondent confidentiality or possibly erase key programs or data files. Third, as general-purpose computers, laptops also are tempting targets of theft, especially when carried into high crime areas. Finally, since laptops must be held in a horizontal position and supported from below when in use, they are not suitable for door-step or through-the-screen interviews.

Portable handwritten character recognition (HCR) devices are being marketed as better solutions to field data collection than laptops. Their plastic membrane cases can tolerate harsher environmental conditions and handling. Their special-purpose design and appearance should discourage both theft and nonsurvey uses. Their note-pad entry units, weighing 3-4 pounds, can be held in one arm like a clipboard while handwritten entries are made with the other. Battery packs and other heavy components can be carried in a shoulder bag and cabled to the note-pad. Because these devices emulate paper-and-pencil data entry, interviewer training is reportedly simple and nonthreatening. These units also are capable of various specialized functions, such as recording tallies, maps, diagrams, and signatures.

Data is entered by marking boxes or by writing characters in defined spaces with

a pen or stylus which the system then converts to ASCII characters. Characters are redisplayed in standard computer characters for confirmation. The same units may later transmit completed interview data over telephone lines to a central office computer.

Two systems demonstrated at the Census Bureau used quite different approaches to character entry and recognition. With the DATAKECH **Scriptwriter** the interviewer uses an ordinary ballpoint pen to complete a paper template inserted over a pressure-sensing plastic surface. The completed paper form becomes a hardcopy backup to the computer-stored data. Only numerals, standard punctuation, and block capitals can be entered, one character per space, and these must be formed in prescribed ways.

The LINUS **Write-Top** uses a special stylus to write on a sensitized glass sheet overlaying an LCD display of the questionnaire or form. When the stylus touches the glass, it is darkened giving a visual sense of writing. Written characters show briefly and then are replaced at about the same point by their recognized equivalent in standard computer characters. Character entry and recognition is user-dependent. Both upper- and lower-case characters are permitted and the interviewer has great latitude in the way he or she forms characters. However, the interviewer must teach his or her assigned **Write-Top** to recognize the way he or she makes each upper-case and each lower-case letter. This takes about 2 to 2.5 hours.

This relatively new technology may have a variety of special uses in survey and census data collection, including block listing, field data entry from hardcopy public records, and monitoring of telephone interviews. At their present stage of development, however, these portable HCR devices do not appear to be suitable alternatives to laptop CAPI for major household surveys. Their display areas are currently too small or too difficult to read; they were not designed to accommodate household rostering or complex multi-page forms; and they do not provide such computer-assisted functions as automatic branching, range and consistency editing, built-in reference screens, and tailored questions based on prior answers. In short, they are primarily data entry devices most appropriate for simple data collection forms.

If the advantages of the HCR devices lie in their ergonomic properties, we may ask whether CAPI hardware can be constructed which has the full screen and computing power of a laptop but the durability, portability, and adaptability for door-step interviews of the HCR devices. The Census Bureau is investigating this option with companies which specialize in the design or manufacture of special-purpose portable computers. For example, although a full-sized screen seems essential for household surveys, a standard keyboard may be unnecessary or a disadvantage. Bryant [36] has suggested that keyboards should be arranged to facilitate one-hand operation by interviewers, especially since few are touch typists.

Survey organizations may find that there is no one best hardware solution for CAPI. Full-sized laptops may be appropriate for long interviews typically conducted by interviewers seated inside respondents' homes. Smaller models with keyboards arranged for one-hand operation may be best for brief, doorstep interviews. Portable handwritten character recognition devices may be best for one-page forms, block listing, and other applications. Rather than seek the ultimate all-purpose CAPI tool, it may become necessary to build an arsenal of hardware alternatives, choosing the most appropriate for each application.

Whatever solutions are found for CAPI's hardware and software requirements, this new technology seems likely to have a major impact on the field operations of household surveys. For example:

1. Field interviewing supervisors and team leaders must be chosen not only for their interpersonal skills but also for their ability to troubleshoot technical problems in CAPI hardware, CAPI software, and in telecommunications.

2. Both initial and survey-specific interviewer training must accommodate new technical topics, such as use and maintenance of CAPI microcomputers, CAPI interviewing, and transmission of assignments and data. At the same time, computer-aided self-instruction packages on these and other topics can be distributed on diskettes or downloaded from headquarters for interviewer home study.

3. Both the mechanics and content of interviewer supervision will change. Batch edits and/or computer browsing of output and call record files will replace visual scanning of completed questionnaire and call records forms. Automatic branching and online edits may eliminate many familiar types of interviewer errors, but supervisors may have to devise new ways of identifying interviewers who require more training in CAPI commands and operations.

4. The check-in and initial review of CAPI interview data transmitted by telecommunications will be performed mainly by computers. The activities of field clerical staff at headquarters or in the regions will largely disappear, while the demand for technical support staff grows.

5. The relationships among headquarters and regional, State, or Province offices (when present) may change with new communication and transmission options. For example, completed interview data might be sent directly to headquarters rather than through local offices, and headquarters may have the capability of sending messages directly to the interviewers.

In short, CAPI may introduce many changes in the way personal visit interviews are assigned, conducted, supervised, checked-in, reviewed, and managed; and these changes will affect the kinds of staff required and the way they are most effectively organized and managed.

2.4 Voice Technology

One indicator of the growing importance of high technology to statistical data collection is the emergence of new sets of advisers to survey organizations. The National Bureau of Standards (NBS) has recently prepared a report for the Census Bureau entitled "Automation of Data Capture for the Census in the Year 2000" [36]. If this report is correct, then CATI and CAPI as we now know them are only transitional modes of data collection.

The NBS report argues that the most natural mode of data collection from human beings is neither paper nor keyboards but speech. It recommends that the Census Bureau consider adding state-of-the-art speech technology to its further development of CATI. Speech technology includes both voice simulation and speech recognition. More sophisticated forms of the latter extend into dialogue analysis and artificial intelligence [43].

The NBS proposed two goals for the use of speech technology in data collection. The first is to "implement fully automated telephone interviews," that is telephone interviews conducted without human interviewers. The second is to develop speech technology as an auxiliary "command/control modality for handheld data entry systems, reducing reliance on keyboard skills."

How far do these options lie in the future? The more one knows about survey field work, and less one knows about speech technology, the more distant (or improbable) these options may seem. They were proposed by the NBS as realistic alternatives for the year 2000 Census, but some early examples already exist. The most promising are for establishment surveys. The Bureau of Labor Statistics, as

noted previously, is already testing a basic form of speech technology, transaction processing, for the dial-in touchtone portion of the Current Employment Statistics Program. BLS also is planning to test voice recognition as a replacement for touchtone entry in the near future.

The use of voice technology in household surveys presents more of a challenge, although early examples can be found in the private sector. Feinberg [12] has described tests of voice digitizing to record the answers to open-ended questions (and the questions themselves if desired) in CATI market research interviews. The respondent's prior approval is, of course, required. The digitized responses may be cross-classified by respondent characteristics and played back for the survey sponsor; or they may be subjected in the future to speech recognition analysis. Stonecash [51] has reported on one of several election polling systems now in use which employ fully automated (interviewerless) telephone interviewing with automated dialing, automated recordings, and touchtone response. The low operating costs of these fully automated systems will undoubtedly encourage their further use and development, but they have not yet attained response rates approaching the standard of most government surveys. The key factor in the use of voice technology for household surveys would seem to be respondent acceptance. Until this new technology is further developed and field tested, its ultimate role in household data collection is not easily predicted and should not be prejudged.

3. Summary and Conclusions

Based on experiences to date, the impact of current and emerging high technologies on survey data collection apparently will continue for decades into the future, perhaps at an accelerating rate. While some survey agencies are still struggling to implement CATI and to understand its full consequences for survey estimates, they must prepare for a much more rapid implementation of CAPI. And while they are just becoming comfortable with the replacement of paper by keyboard modes of data collection, knowledgeable advisers are suggesting that they take the next step into voice technology. Additional options, such as radionavigation systems, computerized self-administered questionnaires, and portable handwritten character recognition, further confound a future offering many potential choices.

If this assessment is correct, then government data collection agencies need to prepare themselves for continued (or at least frequent) technological change in their data collection operations similar to the more familiar periodic upgrades of their data processing systems. Changing data collection technologies may prove more difficult because they so frequently pose organizational and methodological issues. This is readily apparent in government CATI implementation in Phase 1 and may prove at least equally the case for CAPI and voice technology in Phase 2. Rather than viewing data collection technologies as static (or at least rarely changing), it may be necessary to regard them as dynamic. In a world of frequent change, government agencies may find it necessary to develop more efficient and timely methods of identifying, evaluating, comparing, and implementing new data collection technologies as they evolve.

REFERENCES

[1] Bailar, Barbara A., "The Effects of Rotation Group Bias on Estimates from Panel Surveys," Journal of the American Statistical Association, Vol. 70, No. 349, March 1975, pp. 23-30.

[2] Bass, Robert T. and Robert D. Tortora, "A Comparison of Centralized CATI Facilities for An Agricultural Labor Survey," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, forthcoming.

[3] Biemer, Paul, David W. Chapman, and Charles Alexander, "Some Research Issues in Random-Digit Dialing Sampling and Estimation," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 71-86.

[4] Berry, Sandra H. and Diane O'Rourke, "Administrative Designs for Centralized Telephone Survey Centers: Implications of the Transfer to CATI," in Robert M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, pp. 457-474.

[5] Birkett, N.J., "Computer-Aided Personal Interviewing: A New Technique for Data Collection in Epidemiologic Surveys," American Journal of Epidemiology, Vol 127, No. 3, 1988, pp. 684-690.

[6] Bushery, John M. and Lawrence Cahoon, "Labor Force Characteristics," in U.S. Census Bureau, Evaluation of CATI Data Quality and Costs in the Current Population Survey, CATI Research Report No. CPS-2 of the Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, September 1988, pp. 41-49.

[7] Catlin, Gary, Susan Ingram, and Lecily A. Hunter, "The Effect of CATI on Cost and Data Quality," this volume.

[8] Danielsson, L, and P.A. Maarstad, "Statistical Data Collection with Handheld Computers - A Test in Computer Price Index." Unpublished report of Statistics Sweden, Orebro, Sweden, 1982.

[9] Dillman, Don A. and John Tarnai, "Administrative Issues in Mixed Mode Surveys," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, pp. 509-528.

[10] Drew, J. Douglas, G. Hussain Choudhry, and Lecily A. Hunter, "Nonresponse Issues in Government Telephone Surveys," in Robert M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, 233-246.

[11] Dutka, Solomon and Lester Frankel, "Sequential Survey Design Through the Use of Computer Assisted Telephone Interviewing," Proceedings of the Section on Business and Economic Statistics, American Statistical Association, 1980, pp. 73-76.

[12] Feinberg, Barry, "A&S Voice CATI." Paper presented at the American Association for Public Opinion Research Conference, May 1988.

[13] Fink, James C., "CATI's First Decade: The Chilton Experience," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 153-168.

[14] Groves, Robert M. and Robert L. Kahn, Surveys by Telephone, New York, Academic Press, 1979.

[15] Groves, Robert M. and James M. Lepkowski, "Dual Frame, Mixed Mode Survey Designs," Journal of Official Statistics, Vol.1, No. 3, 1985, pp. 263-286.

[16] Groves, Robert M. and James M. Lepkowski, "Cost and Error Modeling for Large-Scale Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 330-357.

[17] Groves, Robert M. and William L. Nicholls II, "The Status of Computer-Assisted Telephone Interviewing: Part II -- Data Quality Issues," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 117-134.

[18] Harding, Dan, Maurice J. Moroney, and Albert D. Frost, "Radionavigation Systems Analysis Evaluation Program for the 1990 Census." Paper presented at the Institute of Navigation National Technical Meeting, January 1984.

[19] House, Carol C., "Questionnaire Design with Computer-Assisted Telephone Interviewing, Journal of Official Statistics, Vol. 1, No. 2, 1985, pp. 209-219.

[20] House, Carol C. and William L. Nicholls II, "Questionnaire Design for CATI: Design Objectives and Methods," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, pp.421-436.

[21] Hubble, David L. and Bruce E. Wilder. "Preliminary Results from the National Crime Survey (NCS) CATI Experiment," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988, forthcoming.

[22] Kefron, J.H. and R.O. Nelson, "Centralized Long-Distance Interviewing -- A Revolution in Data Collection." Paper presented to the Advertising Research Foundation 12th Annual Conference, October 1966.

[23] Kiesler, Sara and Less S. Sproull, "Response Effects in Electronic Surveys," Public Opinion Quarterly, Vol. 50, No. 3, Fall, 1986, pp. 402-413.

[24] Lyberg, Lars, "Administration of Telephone Surveys" in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, pp. 453-456.

[25] Martel, Col. John, "Status of Radionavigation Systems," Proceedings of the Surface Transportation Users Conference on Navigation, Washington D.C., November 1983.

[26] Marquis, Kent and Richard Blass, "Nonsampling Error Considerations in the Design and Operation of Telephone Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 301-329.

[27] Massey, James T., Kent H. Marquis, and Robert D. Tortora, "Methodological Issues Related to Telephone Surveys by Federal Agencies," Proceedings of the Social Statistics Section, American Statistical Association, 1982 pp. 63-72.

[28] Netherlands Central Bureau of Statistics, Automation in Survey Processing, Select Report 4, Voorburg, Netherlands, Central Bureau of Statistics, 1987.

[29] Nelson, R. O., B.L. Peyton, and B.Z. Bortner, "Use of an On-Line Interactive System: Its Effects on the Speed, Accuracy, and Costs of Survey Results." Paper given at the 18th Advertising Research Foundation Conference, New York City, 1972.

[30] Nicholls II, William L, "Experiences with CATI in a Large- Scale Survey," Proceedings of the Survey Research Methods Section, American Statistical Association, 1978, pp. 9-17.

[31] Nicholls II, William L., "Discussion" of papers by Rothschild and Wilson and by Sebestik et al., Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 340-342.

[32] Nicholls II, W.L. and Groves, R.M., "The Status of Computer-Assisted Telephone Interviewing: Part I -- Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 93-115.

[33] Nicholls II, William L. and Carol C. House, "Designing Questionnaires for Computer-Assisted Interviewing," Proceedings of the Bureau of the Census Third Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1987, pp. 95-111.

[34] Nicholls II, William L., Michael E. Haas, and Richard F. Blass, "Cost Models for Computer-Assisted Interviewing in the CPS," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988.

[35] Pafford, Bradley V. and Dick Coulter, "Use of Historical Data in a Current Interview Situation," Proceedings of the Bureau of the Census Third Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1987, pp. 281-298.

[36] Pallett, D.S. (Editor), Automation of Data Capture for the Census in the Year 2000. Final report to the U.S. Bureau of the Census from the National Bureau of Standards under the interagency agreement for fiscal year 1987.

[37] Remington, Patrick L, Meredith Y. Smith, David F. Williamson, Robert F. Anda, Eileen M. Gentry, and Gary C. Hogelin, "Design, Characteristics, and Usefulness of the State-Based Behavioral Risk Factor Surveillance: 1981-87," Public Health Reports, Vol. 103, No. 4, July-August 1988, pp. 366-375.

[38] Rice Jr., Stewart C., Robert A. Wright, and Ben Rowe, "Development of Computer Assisted Personal Interview For the National Health Interview Survey," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988, forthcoming.

[39] Rothschild, Beth B. and Lucy B. Wilson, "Nationwide Food Consumption Survey 1987: A Landmark Personal interview Survey Using Laptop Computers," Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 347-356.

[40] Rustemeyer, Anitra and Arnold Levin, "Report on a Telephone Survey Using Computer Assistance," Unpublished report of the U.S. Department of Commerce, Bureau of the Census, 1977.

[41] Rustemeyer, Anitra, Gerald H. Shure, Miles S. Rogers, and Robert J. Meeker, "Computer-Assisted Telephone Interviewing: Design Considerations," Proceedings of the Survey Research Methods Section, American Statistical Association, 1978, pp. 1-6.

[42] Schwanz, Dennis, Ed Montfort, and John Cannon, "Analysis of Operational Issues: 1987 AHS CATI," CATI Research Report No. AHS-1, U.S. Bureau of the Census, Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, August 1988.

[43] Science, "A Landmark in Speech Recognition" Vol. 240, June 17, 1988, p. 1615.

[44] Sebestik, Jutta, Harvey Zelon, Dale DeWitt, James. M. O'Reilly, and Kevin McGowan, "Initial Experiences with CAPI," Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 357-365.

[45] Sebold, Janice, "Survey Period\Length, Unanswered Numbers, and Nonresponse in Telephone Surveys," in Robert M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, pp. 247-256.

[46] Shanks, J. Merrill and Robert Tortora, "Beyond CATI: Generalized and Distributed Systems for Computer-Assisted Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 358-371.

[47] Herbert Shuldiner, "Computerized Navigator for Your Car," Popular Science, June 1985, pp. 64-67.

[48] Spaeth, Mary A., "Telephone Interviewing Facilities at Survey Research Organizations," Survey Research, Vol. 11, Nos. 3-4, Summer-Fall, 1979, pp.21-25.

[49] Spaeth, Mary A., "CATI Facilities at Survey Research Organizations," Survey Research, Vol. 18, Nos. 3-4, Summer-Fall, 1987, pp. 18-22.

[50] Stock, J. Steven, "How To Improve Samples Based on Telephone Listings," Journal of Advertising Research, Vol. 2, No. 3 (September 1962), pp. 555-564.

[51] Stonecash, Jeffrey, "Automated Dialing, Automated Recordings, and The Effects on Samples." Paper presented at the American Association for Public Opinion Research Conference, May 1988.

[52] Swann, T.C., "Electronic Data Collection in the Petroleum Supply Reporting System." Paper presented at the American Statistical Association Committee on Energy Statistics, April 28-29, 1988.

[53] Thornberry, Owen T. and Gail S. Poe, "NCHS Research on the Telephone Interview: Some Observations," Proceedings of the Social Statistics Section, American Statistical Association, 1982, pp. 296-301.

[54] Tortora, Robert D., "CATI in an Agricultural Statistical Agency," Journal of Official Statistics, Vol. 1, No. 2, 1985, pp. 301-314.

[55] U.S. Census Bureau, Computer Assisted Interviewing: Research, Development, and Implementation. A report of the Central Planning Committee for Computer Assisted Interviewing, 1987.

[56] U.S. Census Bureau, Final Report on Planning for a Post-1990 Redesign of the Current Population Survey. A report of the Oversight Committee of the Joint BLS-Census Committee on The Current Population Survey, April 1988.

[57] U.S. Census Bureau, Evaluation of CATI Data Quality and Costs in the Current Population Survey, CATI Research Report No. CPS-2 of the Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, September 1988.

[58] van Bastelaer, Alois, Frans Kessemakers, and Dirk Sikkel, "Data Collection with Hand-Held Computers: Contributions to Questionnaire Design," Journal of Official Statistics, Vol.4, No. 2, 1988, pp. 141-154.

[59] Waksberg, Joseph, "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Vol. 73, No. 361, March 1976, pp 40-46.

[60] Werking, George; Alan R. Tupek; and Richard Clayton, "CATI and Touchtone Self-Response for Establishment Surveys," Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 273-290.

[61] Werking, George; Richard Clayton, Richard Rosen; and Debbie Winter, "Conversion from Mail to CATI in the Current Employment Statistics Program," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988, forthcoming.

THE EFFECTS OF CATI ON COST AND DATA QUALITY

G. CATLIN, S. INGRAM AND L. HUNTER,

Statistics Canada

1.0 INTRODUCTION

The increased use of CATI in survey organizations over the past few years is due, in part, to expectations that CATI will increase the quality of telephone survey data, reduce the cost of its collection, and improve its timeliness. The early literature on CATI seemed to promise all these benefits but little evidence has been published to indicate that CATI's potential has been achieved.

Previous studies comparing CATI and paper and pencil telephone methods have been published by Coulter (1985), Groves and Mathiowetz (1984), Harlow, Rosenthal and Zeigler (1985), and House (1984). In a two-article series reviewing research on the effects of CATI, Nicholls and Groves (1986) have summarized additional unpublished comparisons and impressionistic evidence. Much of the data that has been presented has been based on surveys with small sample sizes, studies with inadequately controlled experimental designs, or first attempts at CATI by organizations inexperienced with it.

The objective of this study was to evaluate the impact of CATI on survey costs and data quality in a centralized interviewing environment under relatively controlled conditions, and with a study of sufficient size and duration to permit more generalizable results.

The same survey was conducted from a common site by the same interviewers using CATI and paper and pencil methods in alternate weeks. The scale of the test was large. Between February and September 1987, over 10,000 household interviews were conducted, approximately evenly divided between the two methods.

This test was a joint Statistics Canada/U.S. Bureau of the Census (USBC) research project. Statistics Canada used the USBC Census CATI System software and evaluated the impact of CATI using the Canadian Labour Force Survey (LFS) as a vehicle. The LFS is similar to the Current Population Survey conducted by the USBC with questions concerning employment and job search activities relating to a one week reference period.

Initial results of this study on the effects of CATI on data quality and on survey costs during the data collection period have been reported earlier (Catlin and Ingram 1988; Catlin, Ingram and Hunter 1988). Information will be presented comparing response rates, measures of interview quality, survey estimates, cost, interviewer reactions and on the software evaluation. Before turning to these results, we will first review the study's methodology.

2.0 METHODOLOGY

The study took place in the province of Ontario. The target population was similar to that of the regular Labour Force Survey: the civilian non-institutionalized population. Additional exclusions were non-telephone households and collective dwellings with 10 or more people. Non-telephone households constitute approximately two percent of households and collective dwellings less than one percent of households in Ontario (Statistics Canada 1984).

The sample for both methods was selected using random digit dialing (RDD). A systematic sample of working banks was chosen, based on information obtained from Bell Canada and other sources, and a two-digit randomly generated number was added to form a complete telephone number. Screening of the RDD numbers to identify households was conducted using a CATI instrument one month prior to data collection. The households were assigned randomly to either CATI or paper and pencil interviewing and remained in the same interviewing method for as long as they were in the survey.

A sample size of approximately 1,000 households per month was the target for each interviewing method. In both February and March, approximately 333 households (2 rotation groups) were introduced in each method. During each of April and May, 167 households (1 rotation group) were added, bringing each sample size to the required 1,000. From June to September, one sixth of the sample was replaced each month as is the practice in the LFS. A re-interview of one quarter of the September sample of both CATI and paper and pencil treatments was conducted during the last week of September.

Interviewing was conducted from a centralized telephone facility at Statistics Canada's head office in Ottawa. All interviewers worked on both CATI and paper and pencil interviewing in an attempt to reduce the effect of interviewer variability on the test results. Interviewers were trained for a one week period on labour force concepts, and both paper and pencil and CATI procedures. Two supervisors and 8 to 11 interviewers were employed each month.

Two one-week interviewing periods were required each month, one for CATI and one for paper and pencil interviewing. Each month, the interviewing took place in the same two weeks; however, the interviewing methods assigned to each week alternated from month to month. The hours of interviewing were from 8:30 a.m. to 9:30 p.m. on weekdays and as required on Saturday.

The Labour Force interview consists of two parts: the Household Record Docket (also referred to as the control card or household roster) which contains items on household membership and demographics; and the Labour Force questionnaire which is completed for each eligible household member with items concerning labour force activity during the previous week. Although the wording for both interviewing methods was identical to that of the regular LFS, the CATI instrument automatically personalized the wording of the questions based on

answers to previous questions. In addition, for the CATI method, there was automatic branching, enforced probing, editing, and automated call scheduling and sample management. It should be noted that CATI interviews were not conducted using paper and pencil even as back-up in the event of system problems. Respondents were informed that there was a technical problem and that they would be called back as soon as possible.

Both the CATI and paper and pencil samples were processed using the existing LFS system. The edit and imputation, automated industry and occupation coding, and tabulation modules were employed for both methods.

3.0 RESPONSE RATES

CATI may influence response rates either through the ability to contact eligible households for the survey or in its impact on the interviewer-respondent interaction. Response rates were calculated based on all eligible households identified during the RDD sampling in the month prior to their introduction into the survey. Table 1 presents the overall response rates along with t-statistics for the differences between the two methods.

Table 1

Response Rates by Interviewing Method

	CATI	P&P	t-Statistic
Complete	84.3	86.3	-2.05*
Refusal	7.8	8.0	-0.17
No Contact/Absent	7.6	4.8	5.18*
Other	0.3	0.9	-2.83*

* indicates statistical significance at 5% level.

Overall, CATI had a lower response rate than the paper and pencil method. The difference was almost exclusively a factor of less success in contacting households as opposed to refusals. For both methods, 92.1 percent of households responded at least once during their time in the survey indicating that CATI did not produce any more hardcore non-response.

3.1 Refusals

Most of the information in the literature concerning the impact of CATI on refusals is inconclusive. It seems that the interviewers' lack of familiarity and confidence with a CATI system may have some negative impact (Nicholls 1978). However, little data are available on whether CATI increases or decreases the refusal rate in a situation where interviewers are equally adept at both conventional and CATI interviewing. The information in Table 1 shows no significant difference between the two collection methods. Although we speculated that during the critical first few minutes of conducting an interview the more structured CATI interview might provoke more refusals, this did not occur.

Comments by interviewers to respondents that they were entering the information into a computer or the sound of the keyboard had no apparent affect on the respondents' willingness to co-operate. Although there was some month to month variations in the refusal rates, there were no significant differences over the eight months of data. It seems that CATI had no impact on the refusal rate for the survey.

3.2 Household Contact

The literature suggests that call scheduling in a CATI system should improve the ability to contact households by imposing a logic to the frequency and timing of calls during a fixed survey period (Groves and Nicholls 1986). Weeks (1988) states that automated call scheduling can eliminate interviewer error and facilitate the use of sophisticated rules on calling protocols that would be impossible to implement in a traditional manual system.

The call scheduling algorithm for the CATI sample employed (1) the best time to call collected in prior months, (2) the results of unsuccessful calls, (3) the time at which prior calls were made, and (4) appointment information, to assign priorities to calls for a specific time period. The same information was available for paper and pencil interviewing, but no formal attempt was made to prioritize cases. Instead, supervisors assigned a group of sample cases to each interviewer and the ordering of the cases was left to the discretion of the interviewer.

The non-contact component of the non-response included households that were never contacted as well as households that may have been reached but could not complete the interview at that time. As shown in Table 2, the non-contact component was higher for CATI in most months although the difference was not significant in the last three months of the survey. The higher non-contact rate seems to have been the result of several factors: length of interview, computer downtime, and experience and staffing of interviewers.

Table 2

Response Rates and No Contact Rates by Month and Interviewing Method

Month	Response Rates			No Contact Rates		
	CATI	P&P	t-Statistic	CATI	P&P	t-Statistic
February	84.3	84.2	0.04	9.4	5.1	2.30*
March (1)	80.2	87.7	-3.99*	12.9	4.3	5.99*
April	82.7	83.7	-0.58	9.0	5.6	2.75*
May	84.1	87.7	-2.40*	7.9	4.2	3.53*
June	84.6	89.5	-3.27*	5.9	2.9	3.37*
July	83.6	84.7	-0.74	7.8	6.7	0.96
August	84.9	83.8	0.63	6.5	6.5	0.01
September	88.9	87.9	0.68	3.5	3.3	0.20
All Months	**84.3**	**86.3**	**-2.05***	**7.6**	**4.8**	**5.18***

(1) The CATI interview week in the month of March fell on the week of March school break.

* indicates statistical significance at 5% level.

CATI interviews were approximately 20 percent longer than the paper interviews. Therefore, in the most successful interviewing periods, interviewers were able to complete more paper and pencil interviews because they were able to contact more households. For example, Monday interviewing was approximately equally successful for both methods in terms of the number of completed interviews as a percentage of the number of households called, and yet 13 percent more interviews were completed during the paper and pencil interviewing. Later in the week as productivity declined for both methods, there was a larger sample of CATI households remaining to be called. Although the difference in length of interview for the two methods remained fairly constant over time, the average length of interview decreased substantially after the first three months: from 11.90 minutes to 7.95 minutes for CATI, and from 9.99 minutes to 6.27 minutes for paper and pencil. This was primarily a result of the interviewers becoming familiar with the survey and developing confidence in talking with respondents. As the length of the CATI interview decreased to below eight minutes, the number of households able to be contacted during the more productive periods also increased.

Another factor contributing to higher non-contact rates with CATI was computer downtime. Traditionally, Monday to Wednesday were the most productive interviewing days and it was difficult to recover from a loss of calls during this period. Our CATI system ran on a mini-computer shared with other users and there was some downtime in several months. A typical month of interviewing required approximately 300 interviewer hours to complete the sample. One month, we lost four hours of interviewing on a Monday when eight interviewers were on staff; this is a loss of approximately 10 percent of total interviewing time. Numerous calls on a Friday cannot make up for such a loss on Monday.

A final factor contributing to non-response was interviewer staffing. Although the same number of interviewers were scheduled to work the CATI and paper and pencil shifts, over the course of the test interviewers worked fewer hours during the productive shifts with CATI than with paper and pencil. This was caused by interviewer lateness and illness. As with the computer downtime, this loss of interviewer hours during peak periods had an impact on the number of households which could be contacted.

4.0 INTERVIEW QUALITY

The following two sections examine two components of the quality of data produced by CATI and paper and pencil methods. The first deals with item non-response and consistency checks by comparing item error rates. The second compares the codability of responses to a series of open-ended industry and occupation questions.

4.1 Item Non-Response and Consistency Errors

An advantage frequently attributed to CATI is the rigid control of the interview process (Nicholls and Groves 1986). The interviewer is forced through the appropriate question sequence and an entry must be made for each question applicable to the respondent. While this theoretically eliminates errors of interviewers missing questions, it does not ensure that a meaningful answer can be determined by the interviewer and respondent. Although it is possible to prevent entries of "don't know" and "refused" in defining acceptable entries, in practice these responses are allowed in CATI just as with paper and pencil questionnaires. Groves and Mathiowetz (1984) found in a study, using the same questionnaire for CATI and paper and pencil, that the levels of item non-response (combined "don't know", "refused" and "not ascertained") were about the same for a series of demographic and income questions.

The largest effect of the rigid control of the interview process may be in the automatic control of the question sequences. This should be especially true for applications involving complicated branching instructions where the burden of implementing the questionnaire logic is removed from the interviewer. In the same study noted above, Groves and Mathiowetz found that, for a series of 28 questions involving complicated branching, 1.8 percent of the CATI data entries were inconsistent whereas 8.8 percent of the non-CATI entries contained consistency errors.

The LFS processing system edits for validity as well as consistency and an edit report is produced. This report provides the discrepancy or edit failure rate, that is, the percentage of entries inconsistent or blank (includes "don't know" and "refused"). These rates reflect the total error from all sources during collection and data entry. Many sources of error are common to both modes, for example, misunderstanding questions and misrecording answers. However, CATI and paper and pencil are different collection methodologies and some different errors are possible. In CATI, there may be instrument or programming errors while the paper and pencil method has a separate data entry stage where errors may occur. As discussed earlier, the use of the computer in CATI should reduce or eliminate several types of error. Comparing the total error due to collection and data entry illustrates the impact of the different methodologies. Table 3 presents the discrepancy rates by month for the household record docket and the labour force questionnaire.

Table 3

Discrepancy Rates by Interviewing Method

Month	Household Record Docket		Labour Force Questionnaire		Overall	
	CATI	Paper and Pencil	CATI	Paper and Pencil	CATI	Paper and Pencil
February	0.7	1.0	2.0	4.0	1.5	2.7
March	0.5	0.9	2.6	7.7	1.7	4.7
April	0.4	0.7	1.7	6.3	1.2	3.8
May	0.5	0.6	1.8	3.2	1.2	2.0
June	0.8	0.7	1.4	3.6	1.2	2.3
July	0.7	0.6	1.6	3.5	1.2	2.2
August	0.7	0.5	1.4	3.5	1.1	2.1
September	0.7	0.4	1.8	2.9	1.3	1.8
All Months	**0.6**	**0.6**	**1.7**	**4.2**	**1.3**	**2.6**

Overall, the discrepancy rates for the Household Record Docket are identical. For the first four months of the study, CATI error rates were lower than those of paper and pencil whereas in the last half of the study, the paper method had lower error rates. This appears to be a function of differing error rates for subsequent month interviews. For one question in the subsequent month interview, a programming error in the CATI instrument did not ensure that an entry was always made and this is reflected in the Household Record Docket error rates of the latter months. The discrepancy rates for the Household Record Docket also reflect similar numbers of "don't know" and "refused" in the CATI and paper and pencil methods. The Household Record Docket does not have a complicated branching structure. Household membership and demographics are collected in a tabular format for paper and pencil, giving interviewers flexibility in the order in which the data are collected. CATI necessitated the structuring of this portion of the interview, involving, however, only simple question sequencing. As well, although some consistency edits were built into the CATI instrument, the more complex consistency checks (for example, on family composition) were not included.

In contrast, the Labour Force questionnaire has a very complicated branching logic. The difference in the discrepancy rates for this form reflects, for the most part, the effect of the automatic control of the question sequences in CATI. The paper andpencil error rate is double the CATI rate in most months and is three times as large in other months. In one series of questions on the Labour Force form, 23.7 percent of the paper and pencil entries were inconsistent whereas four percent of the CATI data entries were in error. The CATI error rate is not a result of interviewers following inappropriate skip patterns. The discrepancy rates for both interviewing methods also reflect the inability to obtain a meaningful answer.

4.2 Recording Responses to Open Questions

While the control exercised by CATI may lead one to believe that some reduction in response errors is possible, there is concern that other types of error may increase with CATI. Most speculations about increased response errors deal with open-ended questions, where interviewers must type in the respondent's answer to a question (Groves and Nicholls 1986). Errors may arise because interviewers are unable to type the entries quickly enough and, therefore, the entries may contain typographical errors, be incomplete, or unintelligible. Unlike paper and pencil interviews where interviewers often review the responses to open-ended questions for their completeness and legibility once the respondent is no longer on the telephone, in our CATI instrument a complete entry was required at the time of asking the question.

There have been no previous studies which compared textual entries to open-ended questions in the same survey for CATI and paper and pencil. However, a measure of the completeness of recording is available in this study for a series of open-ended industry and occupation questions. The responses to these questions are used in the assignment of industry and occupation codes. As a first step in the process, a computer module attempts to automatically assign the appropriate codes. Table 4 presents the rates of cases automatically coded along with t-statistics for the differences between the two interviewing methods.

Table 4

Coding Rates by Interviewing Method

Coded Automatically	CATI	Paper and Pencil	t-Statistic
February	31.7	29.2	0.88
March	33.5	31.6	0.66
April	29.7	32.4	-0.63
May	34.2	29.9	1.03
June	28.1	35.2	-1.75
July	32.2	33.8	0.35
August	32.9	34.3	-0.32
September	34.0	31.0	0.72
All Months	**32.1**	**31.7**	**0.30**

The rates show no particular pattern. Over the eight months, there are no statistically significant differences at the five percent level and neither method was consistently higher than the other. The number of words used to describe the industry and occupation was also studied. Over the course of the test, the average number of words used was identical for CATI and paper and pencil. It would appear there is no systematic difference in the ability to record entries to open-ended questions.

5.0 SURVEY ESTIMATES

The CATI and paper and pencil data were weighted to allow for the production of population estimates from the survey. In this section, these estimates are compared to estimates produced from the ongoing Labour Force Survey. All LFS data presented are for Ontario and exclude non-telephone households. This study was designed as a controlled comparison of CATI and paper and pencil interviewing and differences between estimates produced by the LFS and either of the test methods may be the result of the test environment rather than the impact of the methodologies. Several important differences between the two test surveys and the ongoing LFS should be noted: the sample size for the CATI and paper and pencil methods was approximately 10% of the size of the LFS; first month interviews were conducted by telephone; interviewers and supervisors were inexperienced; and interviewers did not receive monthly reports on their edit failures and completion rates. The estimates chosen for comparison are household size, unemployment rate, participation rate and the size of occupation and industry groups.

5.1 Household Size

Table 5 presents the average household size by month for CATI, paper and pencil, and the LFS. The monthly differences are small and not statistically significant; however, they do show a consistent pattern over all months. The difference between CATI and the LFS average household size for all months is statistically significant.

Table 5

Average Household Size by Month for CATI, Paper and Pencil, and LFS Telephone Households

Month	CATI	Paper and Pencil	LFS
February	2.26	2.17	2.14
March	2.27	2.15	2.14
April	2.23	2.15	2.13
May	2.20	2.12	2.13
June	2.20	2.17	2.12
July	2.19	2.16	2.12
August	2.20	2.16	2.13
September	2.18	2.14	2.12
All Months	**2.21**	**2.15**	**2.13**

The consistently higher household size in the CATI method may be due to the more structured questioning imposed by the method. Each collection method requires that the interviewers probe for additional household members and persons temporarily away after asking for a

listing of household members. These items often seem redundant to the interviewer. However, the CATI display of these items, requiring an explicit response, likely reduces the tendency of the interviewer to skip these items.

The information on the relationships within households seems to corroborate this explanation. CATI had almost double the number of persons reported with relationships such as parent, grandchild or son- or daughter-in-law compared to the paper and pencil method, and approximately one-quarter more than the LFS. These would be expected to be the types of household members that may be omitted without additional probing.

The size of the difference in average household size between CATI and the regular LFS is approximately three percent which is about half of the estimated under-coverage or slippage rate for the LFS during the period of this evaluation. This may indicate that the more structured questioning with CATI can reduce part of the slippage due to missing household members. Analysis of the re-interview data may shed additional light on this.

5.2 Labour Force Estimates

Two of the most important statistics from the LFS are the unemployment rate and the participation rate. The CATI and paper and pencil estimates of these two rates followed approximately the same pattern as the LFS over the period of the test.The estimates from the two test methods are much more erratic from month to month due, in large part, to their smaller sample sizes.

The unemployment estimates (Figure 1) from the CATI method of collection were consistently higher than the estimates from the paper and pencil method. These differences were significant in May, June and July. The two test methods were each significantly different from the LFS in one month: for CATI, this was May while for paper and pencil, it was June.

The reason for these differences is not apparent. Possible explanations are the more structured questioning and lower error rate with CATI or the higher non-response with CATI or perhaps the difference in the survey week had an impact in these months. The differences between CATI and paper and pencil are likely not due to the interviewers since they worked on both methods, or to the sample which came from the same source and was randomly assigned to the method.

The participation rate estimates from the three sources are compared in Figure 2. There were no significant differences between these estimates. Estimates of the size of major industry and occupation classification groupings were also compared. Again, there were no significant differences between the estimates from the three sources.

Although differences between the three data sources were not significant for most of the survey estimates, some characteristics (unemployment rate and household size), appear to have been affected by the data collection methodology. Plans to incorporate computer-assisted collection into surveys with important data series, such as the LFS, will require close monitoring and further testing.

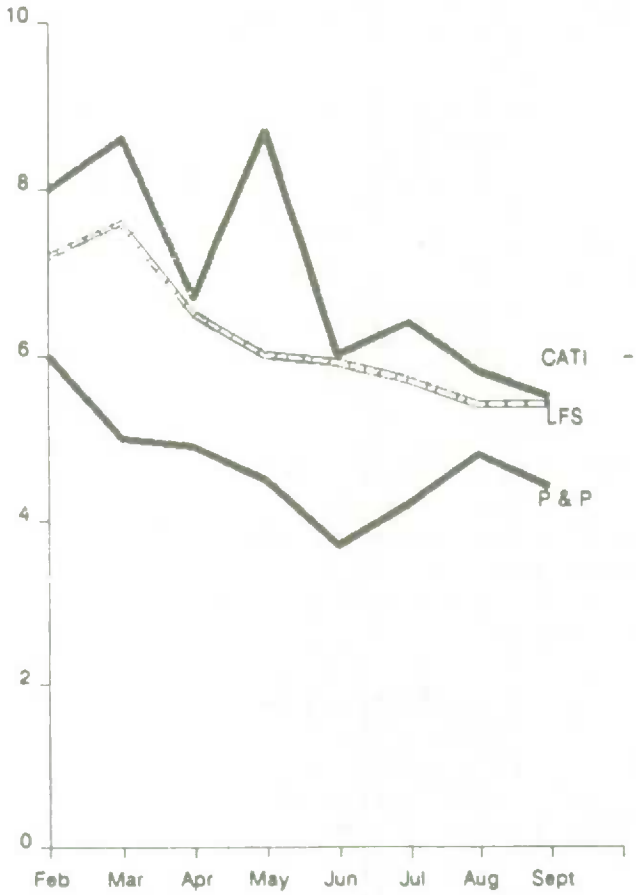


Figure 1
Unemployment Rate by Month for CATI, P & P and LFS Telephone Households

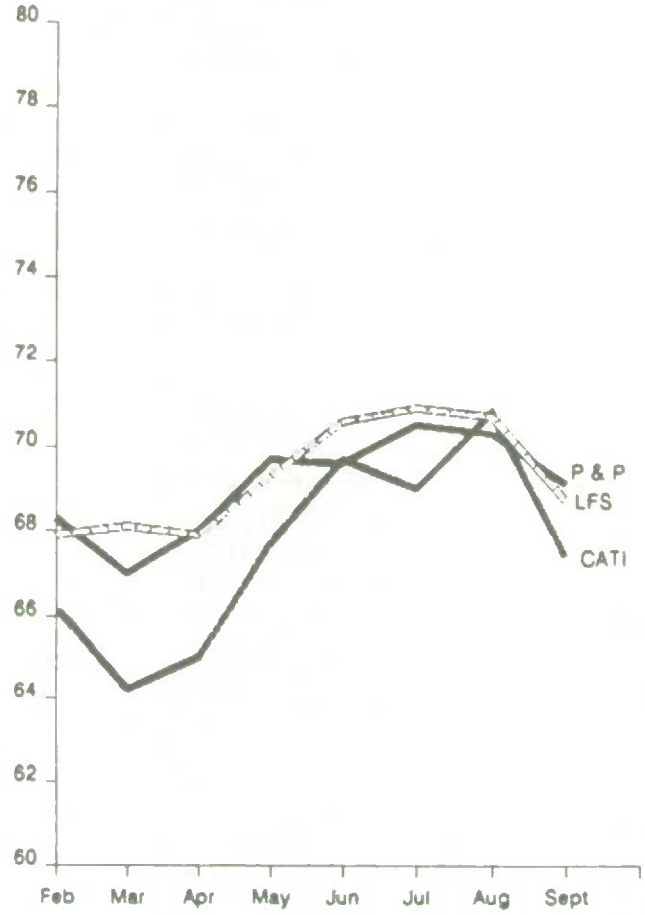


Figure 2
Participation Rate by Month for CATI, P & P and LFS Telephone Households

6.0 COST

At least partial justification for CATI has been that it combines data capture and interviewing and, therefore, should be cheaper. The problem, however, is more complicated than this.

Some researchers speculate that CATI will increase the length and, therefore, the cost of interviews. Harlow, Rosenthal and Ziegler (1985) reported that interviews were approximately 14 percent longer with CATI. Factors noted by other authors include the fact that experienced paper and pencil interviewers begin asking the next question while recording the last and may enter responses to open questions more quickly (Nicholls and Groves, 1985; Groves and Mathiowetz, 1984). On the other hand, CATI decreases paperwork and supervision (Coulter, 1985) with the use of call scheduling. However, the net effect of CATI is not clear.

In this evaluation the costs included are those incurred once the survey infrastructure was present. These include interviewers' and supervisors' salaries, long distance telephone charges, computer costs for the interviewing function of CATI, and data capture for paper and pencil interviewing. The costs were analyzed both as a function of the number of households contacted and the number of interviews completed. As shown in Table 6, the overall differences are small, but CATI appears to be slightly less expensive per household contacted and slightly more expensive per interview completed.

Table 6

Cost Per Unit by Interviewing Method

	CATI	Paper and Pencil	t-Statistic
Interviewers	2.87	2.74	+ 5
Supervision	.63	.73	-16
Data capture/Computer	.54	.81	-33
Telephones*	1.52	1.40	+ 9
Total per household	5.55 (7,076)	5.68 (6,934)	- 2
Total per completed interview	6.63 (5,963)	6.58 (5,986)	+ 1

* Estimates for February, July, August and September.

As suggested in the literature, supervision is less costly in a CATI environment. Supervisors spend less time with the scheduling of calls and the associated handling and control of documents. This was confirmed by debriefing of the supervisors who found their workload during a week of CATI interviewing much less onerous. The relatively small size of each monthly survey may underestimate the saving in supervision that can be achieved.

The information in Table 6 on the interviewer component of the cost is less straight forward. The average length of completed interviews was two minutes longer for CATI. This is consistent with data presented in the literature (Harlow, 1985). The cost for interviewing with CATI and paper and pencil was virtually equal per household contacted despite the longer CATI interview. This is due to the fact that, in the paper and pencil method, interviewers spent more time between calls editing the document they had just completed and determining which household to call next.

A separate data entry cost is eliminated with CATI, however, there are charges for having an interviewer sharing use of a minicomputer in many current CATI applications. When these costs were compared to traditional data entry for paper and pencil, the CATI costs were considerably less. The last component included was telephone charges which were higher as would be expected given the longer interview. The cost per completed interview indicates that CATI is slightly more expensive. The cost per sample household seems to be approximately equal only due to the fact that a larger proportion of the CATI households were non-interviews.

It has been suggested that paper and pencil interviewers can record open-ended responses more quickly than they can be entered in CATI and that this may affect the cost of interviewing (Nicholls, 1978). This did not appear to be the case in this study. Open entries included names of household members and industry and occupation descriptions. In an informal attempt to measure this, we monitored interviews and handwrote responses on a questionnaire as the CATI interviewer entered the same information on a keyboard. It was rarely possible to match the speed of the CATI entry unless the interviewer probed for clarification or spelling. Most of the interviewers hired for the project had some keyboard experience, and this certainly was a factor in their speed of entering information. Others were encouraged to use a micro-computer package that taught typing skills.

We also observed that the interviewers soon became comfortable with the flow of the questions even when using CATI and began asking the next question while entering the answer to the last. This may not occur in surveys which are not repetitive, but it was certainly a factor in this application.

7.0 INTERVIEWERS

The interviewers hired for the CATI project were inexperienced. The staff grew to a maximum of 11 interviewers and 2 supervisors in May and June, 1987, but turnover and reduced requirements resulted in a drop in the staff to 8 interviewers and one supervisor by September. All interviewers were bilingual.

The major responsibility of a telephone interviewer is to develop a professional, co-operative, and friendly relationship with the respondent to elicit the required information. This does not change with CATI; however, the imposition of the technology means that the interviewer must have some additional skills. Much of the information in this section is based on debriefing, monitoring and observing interviewers over the eight months of the survey.

Most of the interviewers had some keyboard or typing skills prior to being hired for the project. Those that did not were noticeably slower during training and required supplementary training to improve their typing skills prior to the beginning of interviewing. After some minimal training, they were able to perform at an acceptable level.

The interviewers must be at ease with the tool that is provided to them for interviewing. Sitting in front of a terminal was a new experience for many of the interviewers. It took several months before these interviewers could concentrate on conducting the interviews without being distracted by what was going to happen when they depressed the next key or what to do when they made an error.

Debriefing sessions were held with the interviewers after the first, third and last month of interviewing. The changing attitude of the interviewers was, perhaps, the most interesting information resulting from these sessions. After the first month of interviewing, 80% of the interviewers preferred the paper and pencil version of the survey. By month three, this had reversed and 80% preferred CATI. Although we expected that once they experienced the advantages of CATI, they would remain converts, the interviewers were slghtly more inclined towards paper and pencil by the end of interviewing. The supervisors were very enthusiastic about CATI throughout the tests. It reduced their work considerably and allowed them to concentrate on monitoring and other tasks.

The interviewers developed an appreciation of some of the advantages of CATI such as being free of paper forms, automatic scheduling of interviews and controlling the questionnaire branching. From the interviewers' point of view, the disadvantages were that it was more difficult to make corrections, nonproductive calls were more time-consuming and a feeling of having less control over the interview situation. The feeling of lack of control was the most often mentioned problem. The interviewers blamed most problems, even those created by themselves, on "the computer".

The feeling of lack of control resulted in the interviewers finding the CATI week more stressful. The reports on hours worked showed that the interviewers worked fewer hours during CATI interviewing. Subsequent discussion with some of the interviewers indicated that they were more likely to not report for an interviewing shift in a CATI week than a paper and pencil week. This higher stress with CATI may ultimately result in a higher turnover of staff with CATI until the staff becomes comfortable with the technology. It may be necessary to have shorter shifts with CATI interviewers or more frequent breaks.

The interviewers became very aware and critical of the CATI system and questionnaire over the course of the test. Response time - the time required for a screen to appear after an answer was entered - was one source of irritation to interviewers. The more adept and quick they became at interviewing, the more instantaneous they wanted the response time to be. Although the response time was fast, one second or less, the interviewers found even this interrupted the flow of the interview. Our response time was restricted, not by the software, but by the terminal transmission rate. The hardware prevented any increase in the rate during production although in testing, we increased the transmission from 4800 to 9600 baud with a noticeable improvement. The response time ideally should accomodate the speediest interviewer's abilities. Interviewers also were critical of screens that did not advance the interview, but only displayed instructions or introductions. Many of these were shortened or dropped during the course of the test to improve the flow of the questionnaire.

8.0 CONCLUSION

For the most part, our data has confirmed the results of other CATI investigations and the optimism that has seen the proliferation of CATI systems in survey research.

The major impact of CATI will be in terms of data quality. In this evaluation a paper questionnaire was translated to a CATI instrument with little alteration. Even in this situation, the item-based error rates were reduced with CATI by 50 percent overall and by as much as two-thirds in the more complex of the two questionnaires included in the survey. In the LFS, each month of collection is independent not relying on previous months' data except as an aid to interviewers. Surveys that employ previous data or edit against historical information will show even greater benefits.

The response rate for CATI in our evaluation was lower. There was no difference in the refusal rate but the non-contacts were higher until the last three months of interviewing. But this would seem to be an impact on cost rather than quality. Interviews were longer with CATI, but as the length of the interview became shorter and interviewers became more experienced, the difference in the non-contact rate was no longer evident. More interviewers may be required due to the slightly longer interview.

Our study has shown that the CATI data collection methodology will likely have some effect on survey estimates. This may be due to the reduction in errors or to the more structured interviewing with CATI. Further testing and close monitoring will be required when CATI is introduced in an ongoing survey program.

The data on cost indicates that interviewing is likely to be somewhat more expensive with CATI due to slightly longer interviews. However, savings will be realized in supervision and the combining of collection and data capture in CATI. The increased quality may result in cost saving in post-survey processing. Additional cost information will have to be collected before final conclusions are reached. The overall impact on individual surveys may depend on their size, duration and the ratio of supervisors to interviewing staff.

In addition to normal interviewing abilities CATI requires the interviewers have some typing or keyboard skills, the minimal skill required can be easily obtained by interviewers who do not already possess it. The interviewers perceived both advantages and disadvantages with CATI. It removed many mundane tasks. However, some actions such as changing answers were more difficult. The supervisors and survey manager found benefits in terms of the survey management and sample control abilities of CATI.

REFERENCES

CATLIN, GARY, and INGRAM, SUSAN (1988), "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing," in Telephone Survey Methodology, Robert G. Groves, et al, (editors), New York: John Wiley and Sons. (Forthcoming)

CATLIN, GARY, INGRAM, SUSAN, and HUNTER, LECILY (1988), "The Effect of CATI on Data Quality: A Comparison of CATI and Paper Methods", Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census. (Forthcoming)

COULTER, RICHARD (1985), "A Comparison of CATI and NONCATI on a Nebraska Hog Survey," Staff Report No. 85, Statistical Research Division, U.S. Department of Agriculture.

GROVES, ROBERT, and MATHIOWETZ, NANCY (1984), "Computer-Assisted Telephone Interviewing: Effects on Interviewers and Respondents," Public Opinion Quarterly, 48(B), 356-359.

GROVES, ROBERT, and NICHOLLS, WILLIAM II (1986), "The Status of Computer-Assisted Telephone Interviewing: Part II - Data Quality Issues," Journal of Official Statistics, 2(2), 117-134.

HARLOW, BERNARD, ROSENTHAL, JEANNE, and ZEIGLER, REGINA (1985), "A Comparison of Computer-Assisted and Hard Copy Telephone Interviewing," American Journal of Epidemiology, 112, 335-340.

HOUSE, CAROL (1984), "Computer-Assisted Telephone Interviewing on Cattle Multiple Fram Survey," Staff Report No. 82, Statistical Research Division, U.S. Department of Agriculture.

NICHOLLS, WILLIAM II (1978), "Experiences with CATI in a Large-Scale Survey," Proceedings of the American Statistical Association, Section on Survey Research Methods, 9-17.

NICHOLLS, WILLIAM II, and GROVES, ROBERT (1986), "The Status of Computer-Assisted Telephone Interviewing: Part I - Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, 2(2), 93-115.

STATISTICS CANADA (1984), "Telephone Ownership in Canada" Unpublished report. Prepared by Gary Catlin, G. Hussain Choudhry, and Hank Hofmann.

WEEKS, MICHAEL (1988), "Call Scheduling with CATI: Current Capabilities and Methods," in Telephone Survey Methodology, Robert G. Groves, et al. (editors), New York: John Wiley and Sons. (Forthcoming)

TWO, NEW, EXPERIMENTAL DATA COLLECTION METHODS AT INSEE

C. BERNARD,

Institut national de la statistique et des études économiques

Two New Experimental Data Collection Techniques at INSEE

C. Bernard

ABSTRACT

The Video Questionnaire

In 1986, INSEE tested a data collection technique for the current monthly industry survey which involved using MINITEL terminals (keyboard and screen) located in business premises. This experiment tested the working capabilities of the new technique, called Video-Questionnaire, and demonstrated its potential for use with surveys having suitable characteristics.

Data Collection Assisted by Portable Micro-computers

Research done in recent years by INSEE in this field culminated in 1987 in a test of the quarterly current survey on a sample of 250 households. The primary objective of this test was to verify that micro-computers were well accepted by survey subjects and survey personnel and that their use was technically feasible.

The main conclusions reached were the following:

- the micro-computer is well accepted by both survey subjects and survey personnel;
- the non-response rate is essentially the same as with conventional surveys, and is not in any case related to the presence of the micro-computer;
- the duration of the interviews is comparable to that for conventional surveys;
- implementation is technically feasible at acceptable cost;
- an upgrading of qualifications is necessary.

The next step will be to carry out a test, in early 1989, with the employment survey prior to its redesign in 1990. This test, involving a sample of 2,000 households, should provide a more accurate indication of the impact of this tool on data quality.

THE IMPACT OF NEW TECHNOLOGIES ON SURVEYS

In terms of new survey data collection technologies, INSEE is working primarily in two directions: video questionnaires and portable micro-computers. Before dealing with the second point, which is the main subject of this paper, I will provide a brief summary of our experiments with the video questionnaire.

VIDEO-QUESTIONNAIRE

Ultimately, this technique should bring about improvements in turnaround time and in the quality of the data collected. In our tests to date, 158 businesses were asked to answer our monthly current industry survey using terminals installed on their premises, for a one-year period. Data for this survey are normally collected by mail. The on-site terminal was a MINITEL, which has a display of 24 lines and 40 columns. It is distributed free of charge by FRANCE TELECOM (formerly PTT) to all telephone subscribers. Its primary function is to replace the printed telephone directory for looking up telephone numbers. The businesses equipped with this device were able, by dialing the access number of an INSEE computer server, to access the query menu and to respond to our survey. The first phase of these tests has now been completed and it has demonstrated that this procedure is technically feasible, although its economic viability remains to be proven. In particular, the anticipated reductions in mailing costs were not realized, since respondents did not faithfully continue with the monthly questionnaires, and a systematic follow-up by mail was necessary. Moreover, follow-up, assistance and some training of respondents still appear to be necessary, although the need for this is gradually diminishing as the MINITEL becomes an integral part of the daily life of the population. The development of such applications is sensitive and poor ergonomic design of the grids causes respondents to give up, and probably also introduces bias in the results.

This technique is best suited to surveys displaying the following characteristics:

- the survey is repetitive;
- a single person can answer all the questions;
- the questionnaire is simple enough to be answered in less than ten minutes;
- there are few grids to be completed;
- the grids are homogeneous in structure;
- the questions are qualitative, since a quantitative question may require research which may not be done with a tool which emphasizes instant response;
- survey respondents are genuinely motivated to answer the survey; and,

- the design of the questionnaire takes into account the characteristics of the tool, which favours a simple, brief and direct form of dialogue.

That is why our second step will involve selecting a survey which meets the criteria described above and whose full-scale implementation will be carried out gradually, so that this new mode of data collection can be observed and consolidated over time.

PORTABLE MICRO-COMPUTERS

Alongside these studies of the Video-Questionnaire, INSEE has conducted research since 1982 in the area of data collection assisted by portable micro-computer. Until 1986, most of our tests encountered technical difficulties, due primarily to hardware and software limitations. The improvement of micro-computer power and reliability in recent years, as well as tests done in this area by other countries, notably the Netherlands, have enabled us to make progress. In December 1987, a test was carried out with the current quarterly survey on a sample of 250 households.

OBJECTIVE

The objective of this test was to measure the technical and human feasibility of such an operation, and to record the reactions of survey personnel and subjects. For the sake of comparison between this new procedure and conventional collection on paper, it was decided to identically reproduce the questionnaire and the manual controls, up to the input phase inclusively.

SURVEY CHARACTERISTICS

The current survey on households contains approximately 240 questions organized into eight sub-questionnaires (composition of household, major appliances, automobile, housing, telephone, general questions, vacations, end of questionnaire). The section devoted to vacations is undoubtedly the most complex: an attempt is made to determine the vacation periods of each member of the household while monitoring overall temporal and geographical consistency.

The sample selected for the test was comprised of 250 dwellings located in a single region of France. The dwellings selected were geographically near those surveyed using the conventional technique. Thus, the same collection conditions (cost, travel, habitat) were reproduced. All the geographical strata were represented in proportions equivalent to their weight in the region. The breakdown of the sample by criteria of social category, size of household and habitat revealed proportions very similar to those recorded in the regional reference sample.

ORGANIZATION OF THE TEST

Ten volunteer survey takers (eight women, two men), who were representative of the survey personnel in the region, participated in the tests. They received two days of training which involved learning how to operate the equipment and putting the final touches on the computer interview program.

A number of the characteristics of the conventional survey were measured before the tests began. We were particularly interested in determining the usual circumstances of non-response, the duration of the survey, the material conditions of the interview (in the stairway in less than 1% of cases, inside the dwelling, etc.).

During the test, the survey takers had to complete a questionnaire following the two days of training, after each interview and at the end of the test. In addition, at the end of each interview the reactions and impressions of households about this new mode of collection were recorded. This information served as the basis for the assessment presented here.

Nine survey takers out of ten unloaded the questionnaires inputted on the premises of INSEE in the middle and at the end of the survey. The tenth computer was equipped with a modem enabling the questionnaires to be transmitted by telephone line.

TECHNICAL IMPLEMENTATION

The machines used in the field consisted of four EPSON PX4s, whose memory capacity had been expanded to 192 KB and six TOSHIBA 1100+ (MS/DOC) with a memory capacity of 640 KB.

The questionnaire was generated in five days on an IPM/PC MS/DOS compatible office micro-computer supporting the SIC (Système Intégré de Collecte) software. This software, still at the prototype stage, was developed by ORSTOM (Office de la Recherche Scientifique et Technique d'Outre-Mer), an agency responsible for conducting surveys in developing countries. The user-friendliness of this product enabled us to develop the entire questionnaire and the required sequencing without major modification. Thus, 12,000 lines of TURBO PASCAL were generated, which once reduced to machine language comprised a module of 31 kilobytes.

ASSESSMENT

It is clear that the objective was fully attained, enabling us to more accurately gauge the impact of this new technology. As a result, we were able to verify:

- that the micro-computer was well accepted by survey subjects and survey personnel.

Indeed, survey subjects showed no particular reaction when they saw the computer. However, greater concentration on and attention to their machines by survey personnel may explain why, in certain cases, the dialogue with the survey subject was less rich, although the survey personnel state that this did not adversely affect the quality of the survey. It should be noted that this concentration was shared by the survey subjects who, according to the survey personnel, thought over their answers longer because they were being recorded on the machine. The

survey personnel adapted well and displayed great enthusiasm for continuing in this direction. It should be noted, however, that they were originally volunteers. They found that the automatic sequencing of questions not only did not bother them, but was quicker and more reliable. They regretted that there were not more controls and experienced difficulties in reading the screen, resulting in visual fatigue. Contrary to what we might think, six survey takers out of ten did not complain about the weight of the equipment, although the TOSHIBA weighed 4.5 kg. However, in our opinion this weight must be reduced, especially if we plan to extend the use of such equipment to surveys such as price surveys, which involve considerable mobility and work primarily in a vertical position.

- that the non-response rate, slightly below that of the conventional survey, is not in any case linked to the presence of the micro-computer.

Results obtained in tests related to primary residences:

	Results from the Test of Dec. 1987	Results from the Conventional Survey of October 1987
Surveys Accepted	92.6%	89.6%
Households not Contacted	3.9%	5.1%
Non-response	3.5%	5.3%

The non-response rate shown above for the test is overall; the low representativeness of the various categories of households in the sample means that the breakdown of non-responses is of little significance. This aspect will be studied more closely during the next test on the employment survey. It should be noted, however, that all the households which refused had not yet seen the micro-computer and that, as in the conventional survey, they did so either on moral (invasion of privacy) or political (economic situation) grounds.

- that the duration of an interview is comparable with the conventional mode.

The interview duration was essentially the same as in the usual survey,i.e. 35 minutes on average. However, survey personnel noted that they conducted fewer surveys per day because of greater fatigue and observed that this had an impact on their remuneration. Because of the automatic locking of the response grids when the machine is turned off, a savings of ten minutes on average was realized with each questionnaire. In the conventional procedure this time is spent at home writing out a clean copy, verifying and completing partial responses by telephone. Six survey takers out of ten reported being hampered by the inability to make such corrections, although no significant difference was observed in the number of partial responses.

- that implementation is technically feasible at acceptable analytical and programming costs.

The use of a questionnaire generator has proven to be a pre-requisite for achieving acceptable development costs. The capabilities of the SIC prototype enabled us to respect the complexity of the questionnaire and to make allowance for contradictory responses by survey subjects when necessary.

However, it must be noted that reproducing the paper procedure as accurately as possible is not the best solution. In particular, the presentation of the vacation and major appliance tables would have been simpler if the capabilities of the equipment had been taken into account. A comparison of the two systems used indicates that:

  - the capabilities of the EPSON PX4 were too limited in terms of screen and memory size for our application and its CP/M operating system was an impediment in our organization. In particular, several return functions and modifications to certain sub-questionnaires could not be installed on this machine.

  - the TOSHIBA 1100+ machine, although very heavy, was better accepted, and was more suited to our expectations; however, its screen is difficult to read.

  - the experiment demonstrated that the machine should be able to run on battery power for at least eight hours.

  Communications between the survey takers' computers and the Institute's computer (an IBM/PC compatible micro-computer) were reliable and easy to implement.

- that an upgrading of qualifications is necessary.

The comments made so far indicate that studies should be undertaken to determine the required upgrading of qualifications and training for survey personnel.

The survey personnel would like to improve their typing dexterity. In this new collection mode, they have more questions about the downstream production process and spontaneously propose supplementary controls.

The experiment highlighted the need to redefine the division of tasks between the field and office staff, notably in the reception, management, inputting and control of questionnaires. From a technical point of view, the office terminal, called the "collection server", should be equipped with electronic mail and file transfer software. It should be noted that, because of the excessively small size of the sample, we cannot conclude significantly, on the basis of a comparison between the data collected and those from the conventional survey, that there is no bias or difference in quality.

The success of this operation has encouraged us to forge ahead by implementing, as part of the preparations for the 1990 redesign of the employment survey, a micro-computer assisted collection experiment which should be carried out in early 1989 on a sample of 2,000 households (50 areas).

The objective of this new test will be to more accurately assess the impact of this tool on data quality, and to implement the collection server function for performing all editing and operational processing.

CONCLUSIONS

The results to date demonstrate that, with this technique, it is possible to collect survey data directly from households on a micro-computer, without risk of rejection. This technique modifies the current organizations and, in particular, requires a new division of tasks between field and office personnel. This latter aspect is perhaps somewhat different from what I have read in the reports on similar experiments in the Netherlands, and this is undoubtedly due to our organization.

This technique is especially well-suited to repetitive surveys characterized by "historical" controls, many screening questions and a limited number of open-ended questions.

In order to take full advantage of the benefits of interactivity and faster turnaround times offered by this too, its inherent constraints must be taken into account, particularly the manner of asking questions, presenting response modalities and verifying them. This may appear to militate against its use in existing repetitive or longitudinal surveys. That is why, taking the weight of the past into account, we are adding an additional verification step which should make it possible to identify any breaks in continuity.

BIBLIOGRAPHY

1. Mission report: 4th Meeting of the Joint Group on General Survey Processing Systems - Statistical Calculations Program, VOORBURG, 8-12 December 1986 - INSEE Software Division, 23 December 1986 - note #1263/640 JPG/MaP. [in French]

2. Report on the Meeting of the Joint Working Group on General Survey Processing Systems - Statistical Calculations Program - London 6-10 April 1987 - INSEE Software Division, note #091/640 JPG/MaP. [in French]

3. Meeting on Statistical Methodology (17-20 February 1986) point 5 - Use of Portable Micro-computers at the Netherlands Central Bureau of Statistics (14 November 1985 CES/AC 48/63) by Micke E. J. BEMELANS-SPORK, Wouter J. KELLER and Dirk SIKKEL. [in French]

4. Working Party on Electronic Data Processing (16-20 February 1987). Micros and local computers at Statistics Sweden by L. NORDBACK - 24 November 1986 (CES/WP 9/229).

5. 35th Plenary Session - 15-19 June 1985. The Impact of the Development of EDP on Statistical Methodology and Survey Techniques - report by Statistics Sweden by Lars LYBERG and Bo SUNDGREN, CES/587 - 1 April 1987.

6. Computer-assisted Data Collection (Netherlands Central Bureau of Statistics - Department of Statistical Methods) 5 December 1986, BPA #15035-86-M1.

7. A Test of the Netherlands Continuous Labour Force Survey with Hand-held AE Computers: Interview Behaviour and Data Quality by VAN BASTEL, Frans KERSSEMAKERS, Dirk SIKKEL, BPA #947-86-M1 - 26 January 1987.

8. The Status of Computer-assisted Telephone Interviewing: Part I - Introduction and Impact on Cost and Timeliness of Survey Data - Part II - Data Quality Issues: William L. NICHOLS II and Robert M. GROVES - Journal of Official Statistics, vol 2, no 2, 1986, pp 93-134.

9. The First Lap-top Computer Fieldwork Project in SSD/OPCS (March 1987)

10. Observations of Prices with Hand-held Computers, Micke BEMELMANS-SPORK, Dirk SIKKEL (September 1984) - BPA #: 13229-84-M1.

11. BLAISE SYSTEM (Netherlands Central Bureau of Statistics) JG. BETHLEHEM, D. DENTENEER, A.J. HUNDEPOOL, M.H. SCHUERHOFF, W.J. KELLER.

    a. The BLAISE System for Computer-assisted Survey Processing - BPA # 14896-86-M5 January 1987;

b. A New Approach to Computer-assisted Survey Processing - BPA # 13436-87-M3 - 6 November 1987;

c. Automating the Data Editing Process with the BLAISE System - BPA # 14263-87-M3 (24 November 1987).

12. Automation in Survey Processing - CBS select 4.

13. Description of the SIC software (Société AGERIS) April 1988. [in French]

STRUCTURE OF THE QUARTERLY CURRENT SURVEY QUESTIONNAIRE

```
IG   General Information  S
                     CM   Composition of Household S
                          MM   Members of Household M
                     GE   Major Appliances S
                          RE   Refrigerator S
                          LL   Washing Machine S
                          TV   Television S
                          LV   Dishwasher S
                          CS   Freezer S
                          RC   Radio/Stereo System S
                          IA   Purchasing Intentions S
                     AU   Automobile S
                          V1   Car M
                     LO   Primary Dwelling S
                     TE   Telephone S
                          US   Telephone Use S
                     GS   General Questions S
                     VA   Vacations S
                          IN   Individuals in the Household M
                               SE   Vacation Taken M
```

CHARACTERISTICS OF THE EQUIPMENT RENTED FOR THE TEST

Two types of equipment were used during the survey:

```
1 -  TOSHIBA 1100+

     Memory capacity: 640 K
     Screen size: 24 lines x 80 characters
     Autonomy: 4 hours
     Weight: 4.5 kg
     Operating system: MS/DOS
```

```
2 -  EPSON PX4

     Memory capacity: 64 K + 128 K extension
     Screen size: 8 lines x 80 characters
     Autonomy: 8 hours
     Weight: 2 kg
     Operating system: CP/M version 2.2
```

Four survey takers worked with the EPSON, while the other six used the TOSHIBA 1100+.

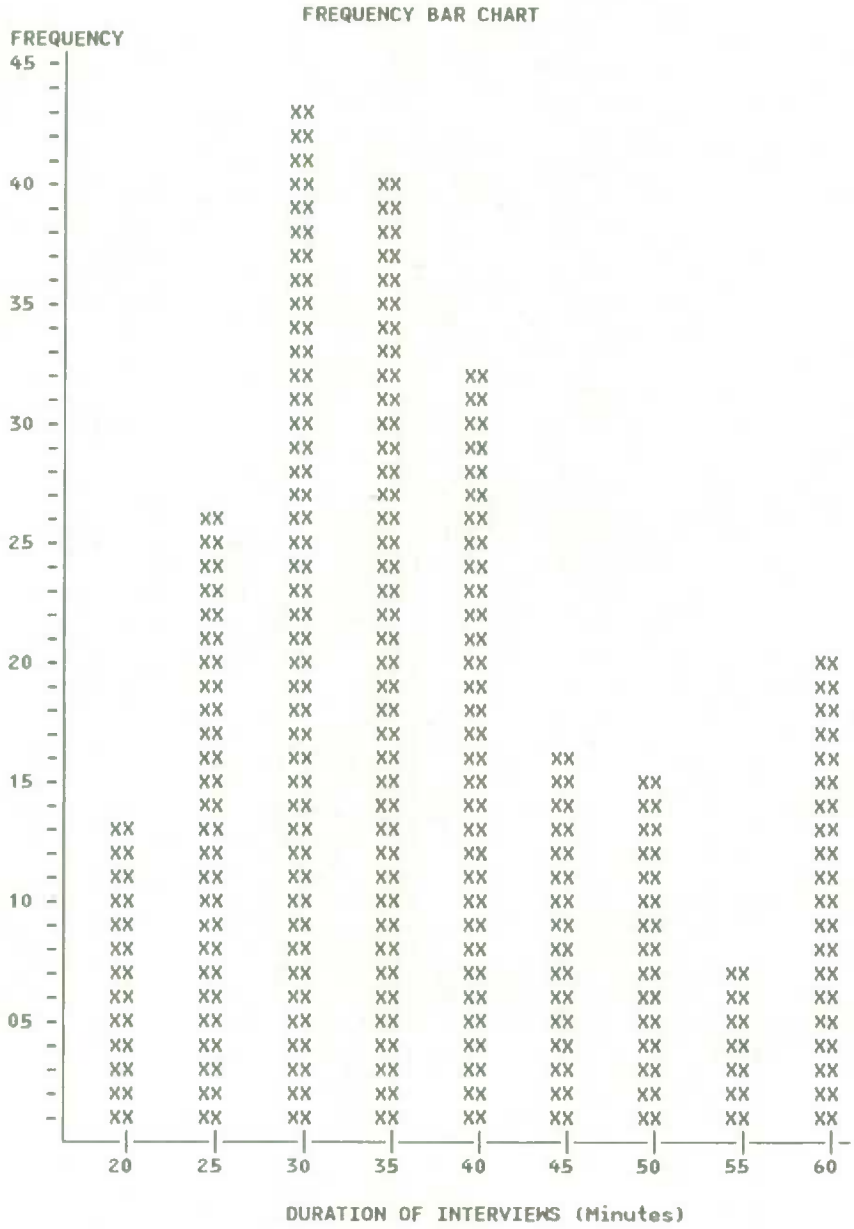
Five malfunctions were recorded, mainly as a result of battery problems.

DATA QUALITY VS. EQUIPMENT USED

A comparative assessment (compared to paper) by survey personnel of the Quality of the Data Collected versus the equipment used.

Quality	Equipment		
Frequency	EPSON	TOSHIBA	Total
Superior	37	24	61
Equivalent	40	97	137
Inferior	5	2	7
Unreported	4	3	7
Total	86	126	212

BREAKDOWN OF INTERVIEWS BY DURATION



AVERAGE DURATION OF THE INTERVIEWS = 38 Minutes

COMPUTER-BASED QUESTION BANKS: WHY AND HOW?

R. P. BAKER,

National Opinion Research Center

INTRODUCTION

The principal goal of every survey is the collection of information about some topic of interest. We pursue this goal by designing data collection instruments that seek a wide range of very specific, well-defined, and naturally interrelated bits of data. By amassing all of these bits of information in one place and exploring the relationships that exist among them, we expect to gain insight into social behavior and attitudes. Expressing this in slightly different terms, we might say that the purpose of a survey is the development of an information base about some important issue.

The survey questionnaire plays a critical role in this information development process. First, and most obviously, the questionnaire defines the questions to be asked. But it does more: it supplements the sample design in determining who our informants will be, and it shapes the research outcome to some degree by expressing our preconceived ideas about how facts and other bits of information interrelate to form knowledge.

Of course, different actors in the survey process view the questionnaire from different perspectives. To the survey questionnaire designer, the questionnaire defines the data the survey is to collect. The designer specifies in precise terms not only the questions to be asked but the range of possible answers. The questionnaire designer translates an analyst's general interest in the marital condition of a population, for example, into specific inquiries about how many people are married, separated, divorced, widowed, never married, and so forth.

To the survey systems designer, the questionnaire specifies metadata, that is, information about the data the survey will collect. These metadata tell us the types of data we can expect, the range of possible values, and the rules by which the data are to be processed. To the systems designer a survey is a process of database development and the survey questionnaire is the dictionary for that database.

Finally, the survey data analyst uses the questionnaire to plan the analysis of the information base. It describes for the analyst the specific information content of the variables he or she analyzes, something about the relationship among them, and the different subpopulations to which the variables may apply.

The focus of this paper is survey questionnaires, their design and management by computer. Its underlying premise is that the survey questionnaire is the key defining document of every survey. With appropriate use of computer technology, questionnaires can be designed with more efficiency, and the consistency, reliability, and useability of survey data can be increased. The paper begins with and returns to question banks, because these both move us toward development of other uses of the computer in survey taking and will be moved beyond their current limits by developments in computer technology and survey research.

The remainder of this paper is divided into four parts. In the first part, some general description of question banks and their uses in surveys is presented. Next, we look at the way in which these uses define requirements for a full-function, computer-based question bank. Part 3 is a look at some key trends in computer technology and in survey research that continue to push us toward the development of complex, multi-function question banks. Finally, a conclusion summarizes the key points from the discussion.

1. QUESTION BANKS AND SURVEYS

The term question bank suggests different things to different people. In its simplest form, a question bank might be little more than a catalogue of survey questions, usually classified by subject matter. The collection might be in machine-readable form and searchable by computer program, or it might be a simple hard-copy document with questions ordered according to topic, perhaps supplemented by a topical cross reference. In its most advanced forms, not yet realized, the question bank may be at the heart of the most complex and powerful mergers of computer technology and survey research.

1.1 Question Banks and Questionnaire Design

Question banks are attractive to questionnaire designers for two reasons. First, one presumes that the questions in them work, that is, they have been written by experts, thoroughly tested, and shown to result in good information when used on a survey. Second, they offer the promise of standardization and thereby increase the likelihood that new surveys will produce data comparable to that collected by previous surveys on the same subject.

A good, recently published example of this simple form of question banking is *The Question Finder* (Owen and Mauro, 1988). This hard-copy reference work has over 5,000 survey questions drawn from a variety of U.S. opinion polls and consumer surveys conducted over the last 25 years. Questions are classified into 91 different subject categories so that questionnaire designers can quickly find questions within their areas of interest.

A frequent expansion on the basic question bank concept is the inclusion of information about surveys on which a question has been used along with the frequency with which respondents chose each answer. An example is the Roper Center's Public Opinion Location Library (POLL), a SPIRES[1]-based data retrieval system with over 85,000 questions asked on a wide variety of surveys including those conducted by Roper, Gallup, Harris, NORC, NBC, ABC/*Washington Post,* and CBS/*New York Times* (Roper Center, 1987). Each entry (question) contains full question and answer text, percentage distributions for the answers, and descriptive information about the survey from which the distribution was drawn including the target population, the number of interviews, the survey organization, beginning and ending dates, and a reference to the specific questionnaire from which the question was taken. Users of POLL can search its database via a subject index, question wording, survey organization, or dates of the survey.

Potential uses of POLL extend beyond those of the questionnaire designer searching for good questions. POLL is both a classic question bank application as well as a valuable information resource for virtually anyone with a substantive interest in public opinion in the United States.

1.2 Other Uses for Question Banks in Surveys

A principal theme of this paper is the degree to which the application of computer technology to question bank development and use is expanding the functionality of the ques-

[1]SPIRES (Stanford Public Information Retrieval System) is a database system for storing free form text and retrieving information based on keyword indexing.

tion bank beyond its use as a menu of choices for the questionnaire designer. Software development projects underway in survey organizations around the world are aimed at expanding the question bank concept to a broad set of potential uses. At least five such uses stand out: (1) automatic generation of questionnaires by computer, (2) evaluation and monitoring of question effectiveness, (3) automated setup of computer-based data capture systems (CATI and CAPI), (4) generation of survey documentation, and (5) data dictionary development to drive extraction software.

Computer-Generated Questionnaires

Imagine an automated system that allows a questionnaire designer to peruse a catalogue of questions, select those he or she wishes to use, and then have the system generate a completely formatted hard-copy questionnaire. Such a system is certainly feasible, especially for relatively short and straightforward questionnaires. However, as questionnaires become more complex the task of building such a system becomes more difficult.

One major area of complication is developing and maintaining a questionnaire's skip pattern. Many of the questions we might encounter in a question bank will have been asked of only a subset of a survey's respondents or in the context of an entire set of related questions. Using them in isolation or even viewing them to judge their appropriateness may be misleading without a clear picture of the context in which they have been used on previous surveys.

This suggests that we must design question banks so that context information is stored along with question text, answers, and response distributions. One means of doing so is to build into the database the capacity to reconstruct the skip patterns of the original questionnaires in which the questions were used. Imagine a user first reviewing a set of questions on a specific topic. The user selects a question that looks interesting and is able to view additional information showing where the question has appeared on different surveys, each with a differing distribution of responses and with different questions preceding it. The interaction of designer and system at this point might look much like that of a CATI interviewer moving back through a questionnaire.The designer might be able to move through as much of the questionnaire--in both directions--as needed, selecting individual questions or even entire sections with their internal skip patterns intact.

As of this writing, developers who work on systems to support surveys are still struggling with ways of representing the relationships among questions in their systems. Designers of CATI systems, for example, must develop ways to track the path of a respondent through a questionnaire so that it is correct according to the questionnaire's skip pattern and can be replicated in the event there is a need to retrace that path to change a previous answer (Nicholls and House, 1987; and Baker and Lefes, 1988). The approaches in use in most CATI systems, while effective for this purpose, seem to have little portability to the quite different problem of constructing questionnaire databases from multiple questionnaires. However, we are beginning to see some applications of graph theory to survey questionnaire design that are very promising. Work being done at the Netherlands Bureau of Statistics (Willenborg, 1987) and at UCLA (Futterman, 1987) is suggesting ways of classifying questions by their location within networks like those created by the skip patterns of complex questionnaires. This work appears to have real application to questionnaire database design, although no such applications have yet appeared.

Question Evaluation and Monitoring

When questions and answers are combined with the response distributions for surveys on which they have been asked, we create a powerful tool for question evaluation and monitoring. One key concern of the questionnaire designer is how well a given question will work. Question banks can operationalize questions so that designers have some sense of the response patterns they are likely to see with a given question structure and set of answers. How much nonresponse does the question generate? Do the responses tend to cluster in ways that will impede analysis? Is the wording subject to misinterpretation by the respondent? Questionnaire pretests are normally used to study these issues and improve the questionnaire before fielding. Well-developed question banks, while not a substitute for thorough pretesting, may be a useful device for early evaluation of questions and even of more pointed and efficient pretests.

Setup of Data Capture Systems

The generation of computer programs to drive automated interviews is still another potential use of the question bank.. The increased use of technologies such as CATI and CAPI has forced us to explore a variety of means for representing questionnaires in forms that can be used by such systems. In most early CATI systems questionnaire designers developed questionnaires in the usual hard-copy format and then a programmer or other technical person translated the questionnaire into a machine-readable format that the CATI system could use to drive an interview. Usually this translation was into a special questionnaire langauge that was in turn translated into a common computer programming language such as FORTRAN. Thus, these systems represented the questionnaire as a computer program. Users of such systems have been known to do some crude question banking by segmenting these programs and retaining blocks of code that handled specific questions or modules in the questionnaire. This is what programmers call reuseable code. It can save a good deal of time in later surveys that use these same questions.

The recent trend in automated data capture system design is away from this approach and toward representing the questions as items in a database or at least as a set of tables, each defining a different item or question. With these systems the questionnaire typically is still developed on paper and then keyed to a computer screen that requests the different components of each question--question text, answer type, valid values, answer text, Boolean conditions for skips, and so forth. Thus, the first step in setting up CATI or CAPI using this design is to create a question bank for the particular survey. Merging databases from a variety of surveys and allowing designers to select questions as well as write new ones seems a logical next step.

As work progresses on CATI and CAPI systems that use questionnaire databases we move closer to use of sophisticated question banks. Both technologies require that we explore more effective ways of representing questions and questionnaires in machine-readable form. Like the generation of hard-copy questionnaires from items selected from question banks, the generation of CATI and CAPI applications requires the development of techniques to represent the relationship among items, in particular, the skip pattern of the questionnaire.

Generation of Documentation and Support Materials

Good survey data requires good documentation to support analysis. Depending on an individual survey organization's standard, this documentation may include a detailed codebook including full question and answer text with frequencies on each question, a record layout, and analysis package setups such as those for SAS or SPSS. The basic raw material for these outputs is the survey questionnaire. Most CATI systems already generate some or all of these products directly from their questionnaire program or file.

Data Dictionaries to Drive Variable Extraction by Analysts

As data base technology continues to mature and its application to the data management problems of survey research broadens, we probably will see a trend toward online, survey databases. The current practice of creating simple rectangular data files with varying levels of documentation imposes a considerable burden on analysts. The problems are particularly acute on surveys using complex questionnaires or multiple instruments, and those featuring longitudinal designs. We frequently hear tales of analysts who spend more of their time merging, reducing, and generally massaging large, complex, multi-file databases than they spend in the actual analysis. The goal of every survey ought to be to provide reliable and pertinent information in a form that analysts can use. For most large surveys conducted in the U.S. this is more dream than reality, although we are beginning to see some new developments that appear very promising (Robbin and David, 1988).

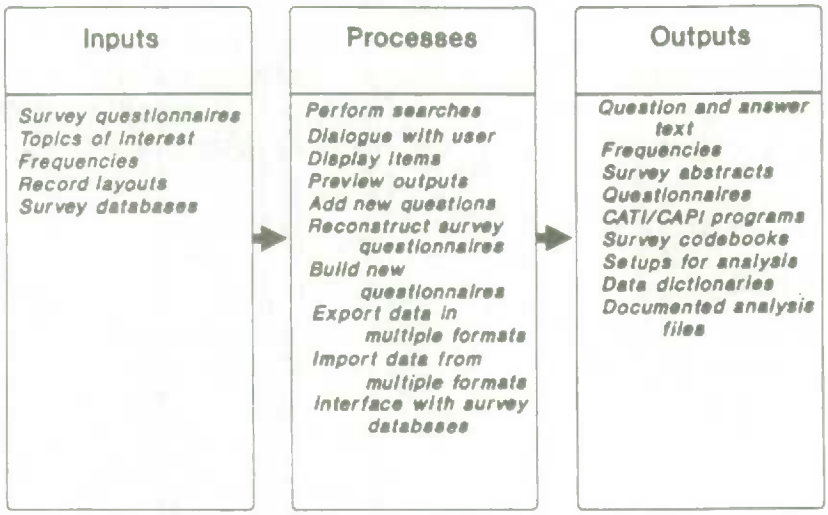
Question banks may be part of the solution to the problem of creating analysis-ready data. Imagine an analyst searching a question bank with the intent of creating an analysis file from a specific survey or even a set of analysis files from several surveys. Once the analyst has selected the questions, he or she has the option of either producing the extract directly or generating retrieval code for an external extract program that runs against the main survey database. The concept here is that of the question bank as a kind of data dictionary (metadata), defining and describing all items in a relational database containing the actual survey data. The question bank itself does not contain the survey data. Rather the question bank is a key component of an overall survey data system directed at the needs of all actors in the survey process--designers, managers, systems analysts, and researchers.

2. DESIGNING COMPUTER-BASED QUESTION BANKS

2.1 System Requirements

Figure 1 is a very general system-requirements model for a full-featured computer-based question bank. Requirements for the system are organized into three components: system inputs, system processes, and system outputs. Systems analysts typically develop such models in the reverse direction of the data flows, that is, they begin with outputs and work back through the processes to the inputs. The principal advantage of this approach to defining system requirements is that the development of one component tells us most of what we need to know to develop the next. Once the desired outputs are known, it is easier to define the system's processes, and, once both the outputs and processes have been described, the inputs are pretty much determined.

Figure 1
System Requirements



The outputs shown in Figure 1 are essentially those discussed thus far. They include complete question and answer text with frequencies, survey abstracts (name, organization, sample size, etc.), time series for questions used on multiple surveys, hard-copy questionnaires, CATI/CAPI code, survey codebooks (with frequencies), setups such as SAS or SPSS commands for analysis programs, data dictionaries for survey databases, and even documented analysis files.

In order to produce this broad range of outputs, the system must be able to perform a number of different operations or processes. It must be able to interact with users to solicit search criteria and display the results. It must be able to accept new questions into the system's database. It must be able to reconstruct complete or partial questionnaires from previous surveys as well as generate new questionnaires from questions selected by the user. The system must be able to generate outputs in a variety of formats--hard-copy and machine-readable--and it must be able to read data in a variety of formats. Finally, if it is to have the capability to create extracts, it must be able to interface with survey databases.

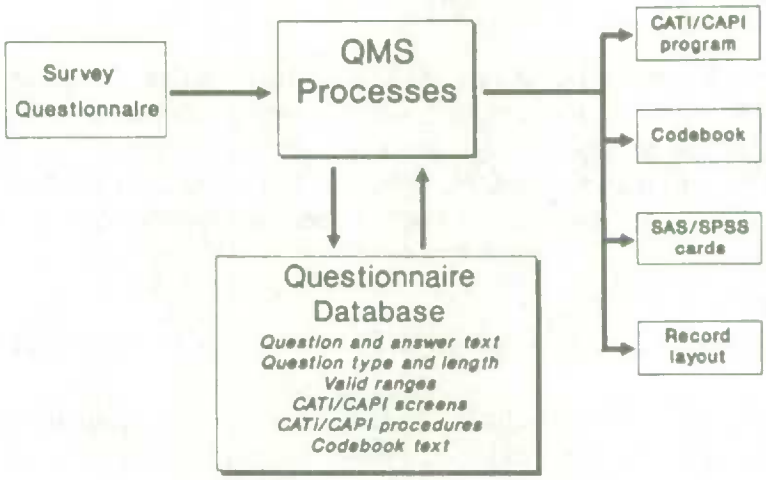
This set of outputs and processes requires five inputs: survey questionnaires, topics of interest to the designer or analyst, frequencies for questionnaire items in the questionnaire database, record layouts describing the structure of survey databases, and, perhaps, the survey databases themselves.

2.2 Recent Work at NORC

In 1985, NORC began a multi-year project to redesign all of its survey production systems (Baker, Beard, and Taylor, 1986). A central feature of the new, integrated system is a questionnaire database with support programs capable of performing some, though certainly not all, of the functions specified for a full-function question bank. The component that manages the questionnaire database is called the Questionnaire Management System (QMS).

QMS is designed around a fairly simple relational data model. Its basic system architecture is shown in Figure 2. In its current conception, the system is used exclusively as a feeder system for other survey production systems. The capacity for question bank development and some interactive questionnaire design is there, but has yet to be developed.

Figure 2
NORC's Questionnaire Management System



The system is interactive, menu-driven, and form-based. It is a multi-user system running on microcomputers on an extensive local area network (LAN) deployed throughout NORC. Under the design currently in production, all questionnaire databases are created on a survey-specific basis. NORC questionnaire designers still develop questionnaires on paper. Once these are complete, staff charged with creating and maintaining the QMS database conduct a careful question-by-question review of the questionnaire with the designer. A questionnaire flowchart is also developed. With the hardcopy questionnaire, question-by-question review materials, and flowchart for reference, QMS users enter question text, answer text, question types and lengths, variable names for questions, valid ranges for answers, and even simple skips onto interactive forms on a PC screen. Forms for alternative question wordings such as codebook text or CATI/CAPI text are also available. The system has a screen painter for free-form development of CATI/CAPI screens and an editor for writing procedures that are used by NORC's CATI/CAPI system to compute variables and indices, manage complex skips, and control other nonstandard movement during computer-driven interviews.

On longitudinal surveys, of which NORC does several, QMS databases from one round are used as the starting point for development of those on follow-ups, a very preliminary form of question banking. Our experience with large-scale modification of these databases is teaching us a great deal about shortcomings in the system and the complexity that is necessary to move QMS up to the level of a true question bank.

Beginning in late 1988, NORC will undertake a redesign of QMS with the intent of broadening its capabilities. One of our design goals is to make the system easier to use so that questionnaire designers can use it directly to set up their own instruments. In addition, we hope to allow for the creation of multi-survey questionnaire databases--that is, to create a

question bank. We also want to improve the system's ability to import questionnaires already in machine-readable form so that they can be read, parsed, and inserted in the database with a minimum of user intervention. The routines that produce CATI/CAPI code will be optimized and the CATI/CAPI procedural language will be made more user friendly.

3. THE FUTURE: DEVELOPMENTS MOVING US TOWARD QUESTION BANKS

Computer technology is changing quickly, and with each new change its applications are broadened. At the same time, research and development in the field of survey research is leading in directions that make it possible for us to realize the benefits of computer technology more effectively than ever before. Whether considered individually or as a group, these developments suggest a continued emphasis on and improvements in designs for computer-based question banks. Five developments or trends, in particular, stand out.

3.1 The Increased Use of Computer Technology in Survey Research

As the program for this conference so clearly demonstrates, computer technology provides the foundation for much of modern survey research. All key steps in the survey process--sampling, data collection, data preparation, and data analysis are managed and accomplished with computers. The development of systems to administer questionnaires and capture data has been especially significant.

CATI is barely 15 years old, yet it has already brought fundamental changes to the way in which we conduct surveys. Its costs are generally lower, its data quality higher than more traditional methods (Groves and Kahn, 1979; and Catlin and Ingram, 1988). Its use is virtually universal in large survey organizations, and it is increasingly used even in smaller ones.

In the last few years we also have seen widespread experimentation with CAPI (Danielsson and Maarstad, 1982; Bemelmans-Spork and Sikkel, 1985; Rothschild and Wilson, 1988; Sebestik, Zelon, DeWitt, O'Reilly, and McGowan, 1988; and Foxon, 1988). By the beginning of the next decade it may well replace paper and pencil in face-to-face interviewing in all of the major survey organizations of the world.

Given this trend toward automated data collection, computer-based question banks are emerging as the ideal tool to design and manage questionnaires for administration by CATI or CAPI. In fact, the practice of designing questionnaires on paper and sending them to technical specialists for set up of automated data capture systems is already giving way to easier-to-use setup systems that can be operated directly by the designer. New systems will be developed that will extend the somewhat primitive design of today's systems to create a multi-purpose tool for questionnaire designers, systems developers, and survey analysts. A key feature of such systems will most certainly be question banking. Experience at NORC suggests that a good first goal is the creation of in-house, organization-wide banks, perhaps each segmented along broad topical lines.

3.2 Improved questionnaire analysis

Full function computer-based question banks require better techniques for representing questions and the relationships that exist among them. The techniques used by most CATI systems, while effective for those applications, are poorly matched to the problems posed by question banks. Current research seems to be focused chiefly on ways to link questions based on the sequence in which they are asked. The applications of graph theory to representation of routing patterns in questionnaires, mentioned earlier in this paper, are extremely promising, and we look forward to the development of systems incorporating such designs.

While the ability to code questions by their place in an elaborate questionnaire skip pattern is an important step forward, we must also be studying ways of relating questionnaire items based on their information content. Current techniques such as assignment of items to subject categories, attachment of keywords, and keyword-in-context searches of questions are almost primitive when compared even to the techniques now in use by designers of sophisticated corporate information systems and data bases. The discipline of information engineering and related techniques such as data modelling (Martin, 1984) focus on the design of relational and network databases in which data items are related through the natural way in which they are combined to form information. Applied to the problem of the question bank designer, these techniques may help us to design questionnaire databases that interrelate items not only by the sequence in which they have been asked of respondents, but by their information content as well.

As an illustration of the difference between these two views of a questionnaire, consider the following question sequence.

Q1. Are you married ?

1. Yes
2. No (Skip to Q5)

Q2. How old is your husband/wife?

Specify:_____

Q3. Is he/she employed?

1. Yes
2. No (Skip to Q5)

Q4. What is his/her occupation?

Specify:_____

Q5. Do you have any children who live with you?

1. Yes
2. No (Skip to Q10)

Figure 3 represents this sequence of questions as they appear to the questionnaire analyst. The questions appear in a specified order with each question defining the next question to be asked. Figure 4 shows this same sequence from the perspective of the data analyst, as a set of data items. Data items are related according to the natural relationships that exist among them. Where the questionnaire analyst sees a series of nodes and dependencies in a network, the researcher sees an entity of interest (spouse) with several attributes (age, employment status, and occupation). Where the questionnaire analyst sees only sequential dependencies, the researcher sees logical independencies. To put it another way, the questionnaire view is of age as precedent to occupation, and, therefore, in some sense age defines occupation. In the data view, age and occupation are related only by virtue of their "ownership" by the common entity called spouse.

Figure 3
The Questionniare View

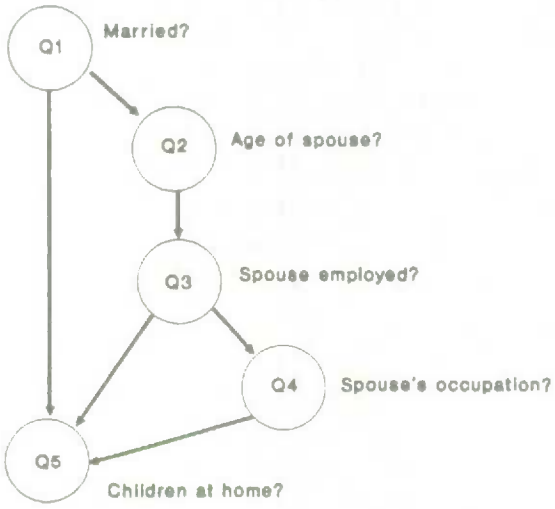
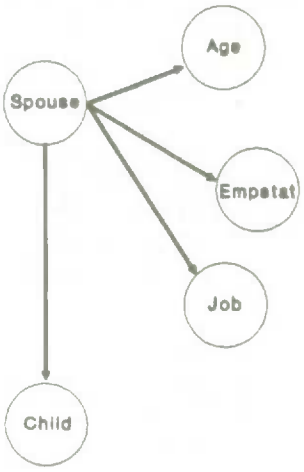


Figure 4
The Data View



3.3 Expert Systems

Information engineering and the design of databases based on the information content of data items leads us to the rapidly developing field of expert systems. An expert system is a computer system designed to duplicate the expertise of a human expert in a particular subject area known as a domain (Edmunds, 1988). A central feature of all expert systems is that they are data-driven, that is, they do what they do by their use of complex data bases called knowledge bases rather than by intricate logic in computer programs. Some well known expert systems include MYCIN--an expert system for physicians to aid in the diagnosis of infectious diseases--and HEARSAY--a prototype system capable of translating natural, spoken language.

Can an expert system be written that duplicates the expertise of a questionnaire designer? The answer is probably, not yet. Certainly there are a number of skills in questionnaire design that could be incorporated into an expert system. They include selecting candidate questions, structuring the instrument into modules, establishing an appropriate skip pattern,

and, of course, formatting the questionnaire either in hard-copy format or for a data capture system such as CAPI. It might even be possible at this point to develop an expert system that could select among alternative sets of answer categories based on the information goals of the designer. Despite these possibilities, the design of a system capable of writing survey questions or even selecting among different question wordings still seems some years away. However, the technology will improve quickly, so survey system designers must begin experimentation soon if we are to take early advantage of the benefits that expert systems offer.

3.4 Hypertext

Hypertext is the generic name given to a new form of electronic document. The defining characteristic of hypertext documents is their reliance on information stored in a network of linked nodes, as opposed to the linear structure of conventional documents (Smith and Weiss, 1988). This network structure allows users to access information in different parts of the document along associative paths rather than with the cumbersome use of indices and tables of contents. Thus, hypertext, or as it is sometimes called, hypermedia, organizes and presents data according to its information content and the interest of the user rather than its physical arrangement. Most current uses of hypertext fall into one of four categories: online documentation for computer systems, dynamic reference works, automated document publishing, or as aids to teaching writing (Barrett, 1987).

It seems obvious that hypertext and its conceptual underpinnings have application to question banks. For example, existing hypertext applications capable of sophisticated text formatting and document construction might be adopted to questionnaire generation. The technology's ability to maintain document databases with elaborate networks of links between items may help us to design questionnaires more or less automatically. Finally, one wonders whether hypermedia techniques might make it possible to design and administer via CATI or CAPI free-flowing questionnaires that follow the respondent's own memory associations rather than the traditional method of plodding relentlessly down the preordained path of the hard-copy questionnaire.

3.5 Research in the Cognitive Sciences

Among the most significant new developments in survey methods over the last several years has been increased interest on the part of survey researchers in the theories of cognitive psychology (Jabine, Straf, Tanur, and Tourangeau, 1984; and Hippler, Schwarz, and Sudman, 1987), particularly those about information processing and their potential application to questionnaire design. The underlying notion of this interest is the belief that a better understanding of the way in which we process information and store it in our memories will help us to design questionnaires that minimize recall error by respondents.

Some recent work by Hastie suggests a way in which this research in the cognitive sciences might be applied to the problems of the question bank designer as well (Hastie, 1987) He describes knowledge and memory structures using knowledge representation techniques developed for artificial intelligence (AI). They include, for example, associative network models, in which semantic concepts are represented as nodes, and relationships among concepts are represented as lines--a format with some similarity to the one used in Figure 4. The artificial intelligence models also include a language that is convenient for representing

operations on network databases, that is, the same techniques that system designers use to represent knowledge bases for expert systems and other AI applications. Thus, the cognitive sciences are not only helping us to design better questionnaires, they are also helping us to see better ways of representing to a computer the information in those questionnaires.

4. CONCLUSION

This paper has sought to broaden our thinking about question banks and their application to survey research. It has argued that question banks have potential beyond their use as an aid to finding standard questions on topics of interest. These extended uses include question evaluation, generation of questionnaires in hard-copy and machine-readable forms, and automated development of metadata for creation of data dictionaries as well as more traditional documentation. In the more distant future, it is possible to imagine questionnaire databases that model a knowledge base to be developed with a survey.

As of this writing, I know of no question bank that meets the broad definition put forth here. Nonetheless, trends in both survey research and in the development of computer technology are moving us toward systems with many if not all of these capabilities. In survey research, we are seeing the continued automation of all steps in the survey process. At the same time, developments in computer technology are moving us away from data processing concepts that until now have guided our automation strategies. The future is one of true information processing with relational and network databases, knowledge representation, and even expert systems.

There probably are several ways to bridge the gap between current computing practices in survey research and new styles of information processing. Among the most promising is recent research by cognitive scientists. Methodological studies seeking to improve surveys through a clearer conception of how people process and store information are pointing survey research toward artificial intelligence as a means both for understanding the structure of knowledge and memory and for representing those structures in a computer.

At the heart of all of this is a conception of survey research as knowledge engineering. In this view, the goal of every survey is the development of a knowledge base in some domain. The survey designer works with analysts and policymakers to identify the information that is needed to study a problem and then develops a knowledge framework to organize it and present it to the analyst or policy maker who will work with it. Surveys have always been designed in this way, so that our basic work remains unchanged. But the conceptual framework and the tools we bring to this work are changing, and will continue to do so, in ways that enhance both the efficiency and the possibilities of survey design, implementation, and analysis.

References

Baker, R. P., "Information Systems in Survey Research," in *Proceedings of the Bureau of the Census Third Annual Research Conference. Hyattsville, MD: 1987.*

Baker, R. P., C. Beard, and J. Taylor, "NISS: The NORC Integrated Survey System," in *Proceedings of the 1985 Public Health Conference on Records and Statistics. Hyattsville, MD: 1986.*

Baker, R. P., and W. L. Lefes, "CATI System Design: A Review of Current Practice," in R. M. Groves et al., eds., *Telephone Survey Methodology.* New York: John Wiley and Sons, forthcoming.

Barrett, E., ed., *Text, Context, and Hypertext.* Cambridge, MA: MIT Press, 1988.

Bemelmans-Spork, Mieke E.J., and Dirk Sikkel, "Observation of Prices With Handheld Computers," *Statistical Journal of the United Nations Economic Commission for Europe,* 1985.

Catlin, G., and S. Ingram, "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing," in R. M. Groves et al., eds., *Telephone Survey Methodology.* New York: John Wiley and Sons, forthcoming.

Edmunds, R. A., *The Prentice Hall Guide to Expert Systems.* Englewood Cliffs, NJ: Prentice Hall, 1987.

Foxon, J., "Field Trials in the Labour Force Survey Lap-Top Computer Project in November, 1987," *Survey Methodology Bulletin,* 23, 12-25.

Futterman, M., "An Approach to CATI Instrument Management Design Utilizing the Properties of Logical Structures," a paper presented at the International Conference on Telephone Survey Methodology. Charlotte, NC: November, 1987.

Groves, R. M., and R. L. Kahn, *Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press, 1979.*

Hastie, R., "Information Processing Theory for the Survey Researcher," in Hippler, H., N. Schwarz, and S. Sudman, eds., *Social Information Processing and Survey Methodology.* New York: Springer-Verlag, 1987.

Hippler, H., N. Schwarz, and S. Sudman, eds., *Social Information Processing and Survey Methodology.* New York: Springer-Verlag, 1987.

Jabine, T., Straf, M., Tanur, J., and Tourangeau, R., *Cognitives Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, DC: National Academy Press, 1984.*

Martin, J., *An Information Systems Manifesto.* Englewood Cliffs, NJ: Prentice Hall, 1984.

Nicholls, W. L., and C. C. House, "Designing Questionnaires for Computer-Assisted Inter-

viewing," in *Proceedings of the Bureau of the Census Third Annual Research Conference.* Baltimore, MD: 1987.

Owen, K. V. N., and J. B. Mauro, *The Question Finder.* Richmond, VA: Smith & Flannery, Inc., 1988.

Robbin, A., and M. David, "SIPP ACCESS: Information Tools Improve Access to National Longitudinal Panel Surveys," *RQ*, 27(4), 499-515.

Roper Center, *User Manual for POLL: The Public Opinion Location Library.* Storrs, CT: Roper Center for Public Opinion Research, 1987.

Rothschild, Beth B., and Lucy B. Wilson, "A Nationwide Food Consumption Survey 1987: A Landmark Personal Interview Survey Using Laptop Computers," in *Proceedings of the Bureau of the Census Fourth Annual Research Conference.* Arlington, VA: 1988.

Sebestik, Jutta; Harvey Zelon, Dale Dewitt, James M. O'Reilly, and Howard McGowan, "Initial Experiences with CAPI," in *Proceedings of the Bureau of the Census Fourth Annual Research Conference.* Arlington, VA: 1988.

Smith, J. B., and S. F. Weiss, "Hypertext," *Communications of the ACM,* 31(7), 816-819.

Willenborg, L. C. R. J., "The Routing Structure of Questionnaires," in *CBS Select 4: Automation in Survey Processing.* Voorburg: Netherlands Central Bureau of Statistics, 1987.

DATA COLLECTION AND CAPTURE FOR SURVEYS

I. S. GREEN AND M. J. WADE,

Statistics Canada

Data Collection and Capture for Surveys

I. S. Green
M. J. Wade

January 3, 1989

Informatics Branch
Statistics Canada
Ottawa K1A 0T6

ABSTRACT

Recent advances in Informatics technology suggest some new alternatives for addressing the significant problems associated with Data Collection and Capture, particularly in a distributed computing environment. These alternatives are explored in the context of recent attempts at Statistics Canada to design and develop generalized software for Data Collection and Capture.

INTRODUCTION

In many instances new technology has already had a profound impact on the way we conduct business at the Bureau. Microprocessors are so commonplace and such an integral component of our daily work activities that their use has become more second-nature, rather than the novelty it once was just a few short years ago. New technology has also made successful implementation of complex systems more probable than was the case in the 60's and 70's. Our *Generalized Edit and Imputation* module serves as evidence.

Some of these technologies have evolved at a startling rate and have significantly altered cost relationships that had prevailed unchallenged for decades. Microprocessor technology has radically changed the relationship between computing power and programmers. Today we most often emphasize optimization of an individual's time rather than that of a processor; a complete reversal of the philosophy we endorsed so strongly in the very recent past.

The emergence of new tools, techniques and dramatic shifts in cost relationships, combined with significant internal changes introduced by the Bureau's Business Survey Redesign Project (BSRP) have brought about a renewed interest in developing a generalized approach to survey processing.

This interest gave rise to the creation of the *Generalized Survey Function Development Team* (GSFD), as part of the overall BSRP. The team has been tasked with providing a suite of systems drawing upon standard methods that address the requirements of the majority of business surveys and, in future, those of social surveys.

Because of its high resource consumption and labour intensive nature, Data Collection and Capture has become the focus of increased attention. This has stimulated a major ***rethinking*** of the ways in which these particular activities are carried-out. As a result of this exploration, several promising opportunities have been discovered.

1. The Current Environment

Presently, Statistics Canada's collection and capture environment is largely reflective of the technologies and associated cost factors that characterized the industry during the mid 70's to early 80's. During that period, as in previous years, the cost of computing power was the dominant factor in determining the location, type and extent of automation. Trends were towards centralization of computing power and resources, and processing was predominantly ***batch*** rather than ***interactive*** in nature; we simply could not afford to let an expensive computer sit idle while awaiting input from an operator.

Minicomputer technology provided a more cost-effective alternative to mainframes for conversion of collected data into an automated format and offered a more productive and versatile approach than key-punch machines. Several were deployed internally and throughout the Bureau's regional offices to initially handle data capture requirements such as those of the Labour Force Survey. Although enhancing productivity, their use was restricted primarily to simple detection of transcription errors in a ***Key-What-You-See*** methodology. This approach did nothing to assist in interaction with the respondent; the key to reducing respondent errors.

This limited automation at the front-end restricted our communication with the respondent primarily to:

- Mail-out/Mail-back of a ***standard*** questionnaire.
- Telephone and transcription to a standard questionnaire for follow-up.
- Transcription from source documents to key-entry forms.

This lack of versatility made collection and capture of data difficult, by the introduction of intermediary steps between the actual collection and the capture. A costly and error prone environment was fostered.

To a significant degree, this front end also dictated what happened in latter stages of data production as we move towards refined data. Even today, in spite of recent initiatives within the Bureau, a typical back end is composed of the following steps:

1. Captured data are transferred from a minicomputer to the mainframe.
2. Batch edits are performed and a ***Turn-Around Edit Failure*** report is produced.
3. Questionnaires are retrieved by Subject-Matter specialists.
4. Corrections are transcribed on turn-around documents and where necessary, follow-up with the respondent is conducted.
5. Updates are recorded using interactive facilities on the mainframe or on a minicomputer.

6. Depending on where the corrections are keyed, steps 1 through 5 or 2 thru 5 are repeated until time no longer permits or it ceases to be cost effective.

This whole approach is characterized by excessive levels of data migration and clerical support. Clearly, by the time respondent errors (also re-introduced transcription errors) have been detected, it is normally too late in time or in the chain of events to obtain a valid response. The only practical recourse is subsequent imputation.

In recent years, the nature of Data Capture and Collection has been changing; this is in response to the Bureau's increased emphasis on *Regionalization*. Interactive translations of traditional mainframe edits have begun appearing on minicomputers, shifting the focus from simple detection of keying errors to the more complex task of identifying those made by the respondent. Although immediate benefits have been realized from this approach, they have not come without cost; proprietary data capture software packages are for the most part incapable of handling many of the new complexities. In response, programmers have resorted to writing large tailor-made extensions to these packages and we now find ourselves responsible for the support of a growing inventory of survey specific applications.

In parallel, the hardware environment for data capture at Statistics Canada has become old and expensive to support; and the operating system and data capture software have not evolved. In short, a replacement is now in order. Essentially, we are faced with two options:

1. Upgrade our existing hardware and migrate the current applications to this platform. In effect we would be *locking ourselves in* to a particular vendor: deja vu.

2. Resist the "quick-fix" and seek less proprietary approaches.

We are pursuing the latter.

2. Communicating With the Respondent

As in other agencies, we, at Statistics Canada continually strive to improve the quality of our data. Ultimately, this requires that a more effective mechanism for communication with the respondent be implemented. Our chosen strategy is to introduce greater flexibility and deal on a more *personal* level.

In view of the various modes of collection and capture that have become widely accepted of late [4], our first inclination was to let the respondent select the method most convenient for him. For those readers that are unfamiliar with these modes of operation, the following brief descriptions are offered.

- *PAPER AND PENCIL INTERVIEW (PAPI)*

  The most familiar and established method. It involves mailing a questionnaire to a sampled member, having him fill in the responses and mail the completed questionnaire back, after which the data is captured.

- *COMPUTER ASSISTED TELEPHONE INTERVIEW (CATI)*

  Although this technology has been known for quite some time, it has not enjoyed wide spread use until recently, as hardware costs have declined. It involves the use of an interactive workstation and a telephone. Reading questions that are displayed on a monitor, a survey operator collects data over a phone and directly enters it into the system. Although this is the basis for CATI, some highly sophisticated implementations also support sample maintenance [5], job scheduling and various reporting capabilities; the United States Bureau of the Census CATI system serves as an excellent example.

- *COMPUTER ASSISTED PERSONAL INTERVIEW (CAPI)*

  With portable microcomputer in hand, an Interviewer makes a personal visit to a respondent's residence or place of work. Here, information can be exchanged in a question and answer session or, if the information resides on documents, it can be directly keyed; thus eliminating the need to transcribe data for later capture.

- *COMPUTER ASSISTED SELF-ADMINISTERED INTERVIEW (CASI)*

  Using this approach, a survey agency supplies data capture software to a respondent; as for the hardware, he has the choice of using his own (if compatible) or equipment supplied by the agency. Under this arrangement, the respondent keys in his data and subsequently transfers the output to the agency.

- *AUTOMATED SOURCES.*

  Today, the vast majority of businesses and government agencies are automated. As a consequence, many are able to provide data in an automated format. Statistics Canada takes advantage of this fact and receives a significant percentage of its data via electronic media such as magnetic tape.

- *OTHER*

  Other technologies for collection and capture of data exist, such as Voice Recognition and Document Image Processing systems. But at the present time they are not sufficiently mature to be incorporated into a generalized collection and capture system and therefore, will not be outlined here.

As an extension to supporting a respondent's personal preference among the various modes for communicating, we can also offer a ***mixed-mode*** of communication. For example, some respondents may wish to communicate via CATI but still insist on receiving a copy of a printed questionnaire. In other cases, a respondent may advise us that, if his completed questionnaire hasn't been received by a given date, to give him a call; in effect, a PAPI/CATI arrangement.

In spite of having accommodated preferences by offering mixed-modes of communication, three other persistent problems still face the respondent:

- *PROBLEM ONE*

  For those surveys that address broadly defined strata, as is the case for Statistics Canada's Census of Manufactures, the result can be a rather extensive questionnaire; large portions of which often may not be applicable to a given respondent.

- *PROBLEM TWO*

  Some respondents are sampled members of more than a single survey and in addition to being subjected to an increased frequency of contact, may also be required to (redundantly) answer questions common to two or more surveys.

- *PROBLEM THREE*

  For those firms that receive a large and/or varied questionnaire, more than one individual may be required to complete it. As a consequence, it may pass serially through many hands before it is actually mailed-back.

We find that to a certain degree, all three problems can be solved by:

- placing an increased emphasis on highlighting the characteristics of a given respondent as a basis for communications, and;

- establishing a sound information model by focusing our attentions on the data we require rather than on cosmetics such as questionnaire structure and phraseology. A data driven rather than say, a question driven model is ultimately more cohesive, providing a basis for edit generation, question selection, elimination of redundant contact and a platform for greater *data accessibility* to the remaining suite of generalized products.

For **problem one,** if *survey X* requires information on a given commodity, and it is known that *Respondent X* manufactures it, then using this information, metadata relevant to this commodity can be referenced. This, in conjunction with our knowledge of his collection preference, allows us to select the appropriate question phraseology (eg. CATI versus PAPI). On the other side of this coin, if it is known that respondent X *does not* manufacture it, the related questions *will not* be selected.

The net result is a greatly simplified and *personalized collection instrument* based primarily on known characteristics of the respondent. The more we know about the respondent, the more helpful we can be.

For **problem two,** being a member of more than one survey is simply another known characteristic of Respondent X. If an *integrated view* of the survey requirements is taken and it is found that they have complementary cycles (eg. all monthlys), then this knowledge can be used to eliminate redundancy and generate a single collection vehicle for all surveys in question.

Problem three requires that the solutions to problems one and two be taken a step further. In those instances, a single, reduced collection vehicle is produced by means of characteristic

selection and integration. To solve this third problem, several unique vehicles must be created. Again, this is another level of characterization. If a total data demand is determined and from this a logical view is produced, it can then be partitioned into categories (accounting data, production data, sales data....) and each generated on separate physical or electronic collection vehicles each destined for one or more individuals within a given business.

Use of characteristics can be further extended to support other techniques that reduce respondent burden and improve the quality of data, such as ***staggered mail-out;*** particularly useful for annual surveys.

As is evident, this technique can be taken to extremes. The challenge here is for Subject Matter and Methodology to determine the practical limitations of integration and characterization as a mechanism for personalized communications.

3. A Design Tool

New technologies have certainly paved the way for sweeping changes and improvements in the ways in which we can gather data from respondents and convert it into automated formats. In the past few decades, dramatic movements have also been experienced in the way systems are developed - each step allowing the designer to concentrate more on solving the problem, simply by reducing the level of awareness required in terms of hardware characteristics, file structures, procedural restrictions and such.

Technologies that allow a designer to approach and solve a given problem by a path most suited to his purpose and style are now available. In addition, these technologies make the designer more aware of the complete set of available facilities, thus promoting greater utilization. As with the other advances in programming, less technical expertise is required.

Using Object Oriented Programming Systems (OOPS), a multi-tasking windows based environment, complete with pop-up/pull-down menu systems and mice, an interface will be created for use by Subject Matter specialists and Methodologists to automate data collection and capture requirements for surveys, with minimal intervention by systems people. Thus, bringing those individuals closer to the actual design of the software.

Through the ***Designer System*** an individual will be able to perform many tasks in any logical order desired. He will be able to define the data requirements of a given survey or collection of surveys, describe the edits that are used to filter them, the questions that are used to acquire them and define the rules that relate a given respondent to those data items, to name but a few of the options.

Several labour saving opportunities exist within this approach, but three are of particular significance:

1. Collection Vehicle Design (Questionnaires, CATI scripts, etc.)

2. Edit Specification

3. Screen Design

Compared to conventional methods, using a computer assisted approach to design the physical layout of a questionnaire can be a highly productive venture, even when using a character-based system, as was clearly demonstrated by the generalized prototype developed at STC in 1987. In this approach, standardization began at the question level; questions were placed in logical groupings called blocks which were used to construct ***personalized*** questionnaires.

To describe a question, the designer was given up to three lines of 130 characters each to accept the actual text. The monitor only supported the standard 80 characters per line, so if the text exceeded that limitation, horizontal scrolling was used to bring it into view. At no time however, could the entire question be seen if it exceeded that limit.

The print image (font size and type) for individual words within a question could be altered by nesting special characters within the text. These served as signals to a printer driver.

To describe a block (a grouping of logically related questions), the designer was provided a mechanism to enter supportive text in much the same fashion as in defining a question. Following this, special codes were entered which defined the member questions, indicated where a blank line should be printed, whether a dividing line should be printed, and so forth.

To determine what a question, block of questions or a questionnaire would actually look like, a special print function was provided in another portion of the system to generate the image on a laser printer. Using this design tool, a questionnaire could be described and printed in a session lasting only a matter of hours; an impressive step beyond conventional methods.

Using a graphics and windows based, multi-tasking system combined with a large screen monitor, we can create the same result as above while reducing the time and effort required by yet ***another order of magnitude***.

Using graphics capabilities and a large monitor, questions can be keyed directly onto a blank screen or pallet and viewed in their entirety, whether they exceed 80 characters in length or not, because the designer can ***zoom in or out*** to make text fit the monitor. Nesting of special characters to change font sizes and styles for individual words within question text would be unnecessary, because by using function keys (or some other creative alternative) dynamic presentation of fonts could be effected right in front of the designer's eyes - a true ***What-You-See-Is-What-You-Get*** (WYSIWYG) environment.

If we employ Object Oriented Programming and define questionnaires, data capture screens, edits and say, security as ***objects***, then a great degree of flexibility in a design session can be realized;[1] especially if it is utilized in a multi-tasking environment. For example, if, after a question has been defined, a designer wishes to continue concentrating on questionnaire design and immediatedly describe another question in a block, he can do so; or he can perform whatever activity or activities serve his purpose at the time.

In addition to allowing a designer a great deal of flexibility in the order in which he carries out his activities, the new programming techniques also allow for simulation of ***how we work at our desks***. In effect, they provide an electronic desk top. Perhaps this point will become clear as we continue through the example questionnaire design session.

---

[1] Partitionning design into several discrete elements will require that an integrity checker be built to ensure that all necessary components exist before moving to production.

After having keyed the text for all questions intended for membership within a block (assuming that is the designer's intent), the work can be saved, the image reduced to an icon and moved to another area of the screen, much like an individual would move a piece of paper to another area of a desk top.

Next, the designer could move to ***block creation mode*** and begin keying the supportive block text using the same facilities used to design questions. Following this, instead of indicating to the system by use of codes that a dividing line should be drawn on the questionnaire to delimit the block, he can simply ***draw it;*** akin to using a ruler and a pencil.

To associate questions to the block, as an alternative to providing a series of identification codes, he can retrieve the reduced image of the questions in the other part of the screen (much like an individual would grasp the sheet of paper in another area of the desk top), adjust it to the proper size (or use scissors to cut to size) and insert it into the block image (glue the question images into the block image). In effect, a graphical ***cut-and-paste*** operation.

At this point, no printing of the block is necessary; he can see exactly what it would look like on printed material just by looking at the image on the screen. Again, he can print if he wants, but ***no procedure is enforced.***

In terms of physical image, data capture screens are conceptually the same as questionnaires, except that one is mapped to a sheet of paper and the other to a monitor. Both have text and graphic characters, so many of the facilities and techniques used to compose questionnaires can also be used as labour saving mechanisms to paint screens.

During the development of the generalized data collection and capture prototype, we were able to dynamically paint ***default personalized screens*** as long as two items of information were defined:

1. the data items belonging to a survey.

2. the relationship of the respondent to the data (ie. data demand).

Using this information, the capture system was able to derive prompts and determine where these should be placed on a screen in a useful manner while adhering to the order in which they were asked of the respondent. Additionally, it could apply default edits and generate default ***context sensitive help*** for each and every field.

The point being made here is this:

If the image of a data capture screen is not conceptually different from that of a questionnaire, and if we have information about survey data, including how it is related to a given respondent, then it follows that we can also dynamically generate default ***personalized questionnaires.***

As with painting data capture screens, limitations will exist. For example, a designer may wish to make some aesthetic adjustments to a generated image. This concept does however, warrant further consideration.

A third major activity which provides opportunities for significant reduction in required levels of effort and expertise, is edit specification. Some excellent approaches to this problem are demonstrated in commercial systems.

In addition to their screen painting capabilities, many suppliers of microcomputer-based data capture software include creative edit selection/definition facilities that extend their usefulness beyond that of simple key-punch substitutes.

Some of the more successful vendors have addressed the problem by building libraries of generalized edit routines and supplied facilities such as context sensitive help (messages to be displayed as part of actions to failed edits). These edits then, can be selected in an *off-the-shelf* manner by their customers' capture system designers. With a degree of difficulty ranging from a *fill-in-the-blanks* session for parameter driven edits to a *Yes-or-No* session for optional edits, the need for programming expertise and tailor-made products has been significantly reduced.

Driven to improve their product's value, certain vendors have pursued higher degrees of functionality by developing a *data capture language* which their customers can quickly learn and use to build edits not included in the generalized libraries. These languages support sequential execution, branching, iteration, several levels of nesting and can be referenced from many places within an application as subroutines; in effect, third generation languages.

Although in many ways the equivalent of tailor-made extensions, these data capture languages offer the advantages of being relatively easy to learn, they greatly extend the versatility of a capture package, and the designer is not required to purchase or become familiar with a compiler and linker.

These applications do however, fall short in two areas:

- First, by necessity their focus has been on private industry. A huge portion of their market is composed of retailers, banks and so forth. As a consequence, they are geared to satisfying the edits required to capture a VISA slip; a somewhat less formidable object than a Census questionnaire.

- Second, one can safely assume that on many occasions, edits defined by data capture languages will be quite extensive. Although these languages are compiled, assistance for the author is limited primarily to simple syntactical checks. Problems in logic are encountered at run-time and for the most part would require a *systems person* to resolve them.

We can solve the first problem by investigating some of our tailor-made capture programs and mainframe edit programs. It is quite probable that a large number of edits peculiar to statistical agencies can be identified, generalized and added to the list of off-the-shelf products.

As for the second problem, *lexical analysers* can be purchased or built if necessary, that check a designer's edit rules for completeness, redundancy, and for conflict; thus, further reducing the need for programming expertise.

4. The Next Environment

Statistics Canada offers several different computing environments. Ranging from mainframes to minis to micros to hybrid technologies, in a mixture of networks, it spans thousands of kilometers and several time zones - a truly distributed configuration.

Although collection and capture activities can occur virtually anywhere within this framework, they are concentrated primarily within two specific organizational units: Headquarters Operations Division (HOD) and Survey Operations Division (SOD). While having central representation, SOD's computing environment is geographically partitioned into eight regional offices and associated satellites. HOD has assumed responsibility for collection and capture of data for those surveys that can be more effectively serviced through a centralized approach. Both environments are predominantly minicomputer based.

Although the impact of the Bureau's new approach to data collection and capture is expected to be widespread, it is these two organizational units, along with their respective clients that will reap the greatest benefit.

When implemented, it is anticipated that in the order of two hundred or more surveys will be supported, all contributing to a highly active and varied environment.

Perhaps the best way to appreciate what is being proposed here, is to describe a hypothetical monthly survey and walk through one of its cycles. Before this can be done however, the following four assumptions are made.

Assumption 1. Survey X is a business survey and a central frame or master list of in-scope Canadian businesses exists. From this list, all samples are drawn and respondents that are sampled members of two or more surveys can be identified.

Assumption 2. Through a Designer System, Subject Matter specialists and Methodologists have described the survey's requirements. These specifications, in the form of questions, capture screens, selection criteria, edits and help messages have been thoroughly tested through the prototyping facilities of the designer system.

Assumption 3. Survey X has already been initialized. All specifications (rules) have been ported to the work site(s) and loaded into the production system.

Assumption 4. Survey X's list of collection entities are defined each cycle by a system external to collection and capture. A collection entity can be regarded as a logical record describing a given respondent's characteristics in terms of the surveys in which he is sampled, his communication preferences, contact name(s) and address(es), follow-up date(s) and information indicating data that is and data that is not required of him.

To begin a cycle, a work site's case management system is initialized. Each contact (there may be more than one) within any given business is treated as a separate case and is assigned its own unique identification code.

During this update, each case is analysed.

For those that prefer to communicate by questionnaire (PAPI), their characteristics are examined. Comparing individual characteristics within a case against selection criteria, a data demand is determined. Using this data demand to reference a ***standard question bank***, questions and supportive text (such as instructions, logos, dates) are selected and a personalized questionnaire with its unique case identification code is generated on a laser printer, complete with

mailing instructions. The case management system is updated to reflect this status.[2]

For those cases preferring computer assisted personal interviews (CAPI), the case information is targetted to an interviewer.

Upon acceptance of responsibility for the cases by an interviewer, they are transferred to a floppy diskette for use in his portable microcomputer. Additionally, all of the edit rules, help messages, and reference tables associated with survey X are transferred to his CAPI system shell. The case management system is updated to reflect the *assigned status.*

For those respondents who prefer to capture the data themselves (CASI) or extract it from their own computer systems (automated sources), a *reminder notice* complete with names, address information and case identification codes is generated . The case management system is updated to reflect the printed and mailed status in the same manner as was done for questionnaires.

The remaining cases have selected CATI as their preferred method of communication and are placed on a CATI queue. As in all other modes, the case management system is updated accordingly.

It is at this point that initial contacts for all of survey X's cases have been effected or put in motion. The focus of attention shifts from the management of information moving out of the work site to management of information flowing in.

For CATI, teams of interviewers are identified and the scheduler begins drawing cases from the CATI queue(s); allocating them to individuals specifically assigned to a given case or to an interviewer who is available. To conduct a session, the characteristics of each case are compared against selection criteria to create a data demand in much the same way as was done to produce a personalized questionnaire. In this instance however, when referencing the standard question bank, the CATI version of the questions are drawn rather than PAPI.

At the same time, completed questionnaires are being received from the field. As they arrive, clerks scan the case id numbers with OCR devices, and the status information is posted. Concurrently, the case management system assigns groups (batches) of questionnaires to individual folders, which are in turn assigned to physical storage slots.

As the site manager's capture operators become available, they are given folders for survey X. Before the contained questionnaires are captured, the folder id numbers are scanned and registered as *transferred to capture.*

After completing a folder, a capture operator indicates *folder complete* to the document control system. It then compares the questionnaires keyed against those registered to the folder and notifies the operator or supervisor of any discrepancies. Upon resolution, the folder is returned to its registered filing slot and the document control system (within case management) is updated.

Questionnaires considered to be incomplete or as having respondent errors then have their corresponding cases added to a CATI follow-up queue.

---

[2] Anticipated volumes suggest that high speed central printing facilities will be used.

Self administered and automated source cases being returned from the field are automatically *checked in*. All cases of these types are subjected to a batch edit. They are then treated in the same manner as a paper questionnaire.

As cases are activated for follow-up at a work site, interviewers can:

- prioritise and pursue the outstanding balance of data within a case, or;
- select only those cases satisfying a predetermined selection criteria (relative weight of the unit within the sample), or;
- adhere to a ***minimum compliance rule***, by pursuing only those responses satisfying a predetermined selection criteria (eg. relative importance of a data item within a questionnaire), or;
- pursue cases based on combinations of the above strategies.

In essence, this covers the main ingredients of activity at any given work site. It should be noted however, that the system will support movement of cases ***between*** work sites.

What must be clear to the reader at this point is:

- the high degree of automation and communication involved.
- the flexibility of the environment: mixed modes, generation of personalized questionnaires and CATI scripts, etc.
- that case management can assume a global role over all supported modes of collection and capture.
- that document management in conjunction with case management and the variety of available mechanisms for collection and capture provide the site manager with a very high degree of flexibility and control.
- that the automated elements of the proposed production system are highly integrated.

5. *The Manager's View*

It should come as no surprise that developers of commercial data capture software had targets other than statistical agencies in mind when they created their wares. In many cases, they sought to maximize their product's marketability by appealing to a broadly defined client base. It is this aspect that makes certain of their features to be of minimal use; particularly the ***management information*** facilities.

Through examination of the proprietary minicomputer packages currently used within the Bureau, along with the several microcomputer based systems now available, we find that site managers are by and large, hand-cuffed by the vendors to a myopic view of key-strokes, batches and primitive forms of quality control. To address the latter problem, the Bureau recently contracted the development of a sample verification system for use on Honeywell minicomputers, as an alternative form of quality control.

The larger problem still remains.

It can be significantly reduced however, through implementation of a truly generalized data collection and capture system. By integrating the suite of feasible modes, together with a global approach to case management, managers can expand their view of the processes simply by using each mode as a benchmark against the others. If we knit a comprehensive document control system into case management, still more of the production process can be monitored. Beyond this, by building upon a relational platform, definition of standard and ad-hoc indicators is possible. Collectively, these aspects will allow managers to enjoy a far greater degree of flexibility in many aspects of monitoring and control over collection and capture processes.

To offer a few examples:

1. If a production system's data model supports combination of survey demands into a single collection instrument, then it follows that some forms of cross-survey analysis can begin in the earliest stages of data production.

2. If a site (or survey) manager can partition his respondents into those that respond in a timely fashion via PAPI and those that do not, then the latter can be switched to CATI. This has potential to speed the flow of data and allows a manager greater control over resource utilization.

3. If the production system can be regarded as being composed of "n" work sites rather than each work site having a production system, then management has opportunity to take a higher level view of the collection and capture process.

Site managers, Subject Matter and Methodology are faced with a difficult challenge in determining how best to make use of this promising opportunity.

6. A Look To The Future

Now, more than at any time in the past, management is placing emphasis on ways to reduce software development and support costs. As part of our long term strategy in pursuit of this goal, we will:

- *ADHERE TO PORTABILITY STANDARDS*

  By making use of systems developed in the "C" programming language and using it for in-house development. It is our hope that as our hardware platforms change, the suite of generalized data collection and capture software and associated applications can be migrated in a relatively seamless fashion, avoiding the substantial conversion costs incurred in the past.

- *MAKE INCREASING USE OF GENERALIZED SYSTEMS*

  Using this approach, Subject Matter and Methodology are drawn closer to the actual software design, reducing the need for a costly middle man (systems person). As there are fewer people required to communicate requirements, a higher quality finished product can be obtained. In addition, by having all applications run under the umbrella of *one produc-*

tion system, only one systems team is required for maintenance. These specialists can then provide a higher level of support to their user community.

- *ADHERE TO INTERNATIONAL STANDARDS*

  By making use of data management software (RDBMS) and programming languages (SQL) recognized and supported as standards by the international statistical community, opportunities for exchange of software, expertise and other forms of cooperation will arise; thus, further reducing the cost of software.

Looking still further into the future, these generalized systems are in effect ***an intelligent application of software*** which, among other things, deal interactively with people in the design and creation of survey applications. As such, they serve as a rich source of the rules and facts needed to create the knowledge bases for the next generation of survey processing software; expert systems.

BIBLIOGRAPHY

1. Doucet, E. (1988) *Redefining Software Portability; A new Design Objective For Applications* Research And General Systems Subdivision, ISD, Informatics Branch, Statistics Canada.

2. Emery, D. (1987) *A Strategic Review Of Data Capture* **(Internal Document)** Research And General Systems Subdivision, ISD, Informatics Branch, Statistics Canada.

3. House, C.C. (1985) *Questionnaire Design With Computer-Assisted Telephone Interviewing* Journal Of Official Statistics, 1(2). pp.209-219

4. Lyberg, H. (1985) *Plans For Computer-Assisted Data Collection At Statistics Sweden* Proceedings of the 45th Session. Bulletin of the International Statistical Institute, Book III, Topic 18.2, pp.1-11.

5. Nicholls, W. (1983) *Development Of CATI At The U.S. Census Bureau.* Proceedings of the American Statistical Association, Section on Survey Research Methods. Washington, D.C.

6. Shanks, J.M., Lavender, G.A., and Nichols, W.L. (1980), *Continuity And Change In Computer-Assisted Surveys - Development Of Berkeley SRC CATI.* Proceedings of the American Statistical Association, Section on Survey Research Methods, pp.507-512.

7. Von der Heyde, C.A. (1987) *Infrascope - The State Of The Art Of CATI In Germany.* Infratest Forschung, Stechprobenanlage Und Statistik, West Germany

COMPUTER ASSISTED SURVEY PROCESSING ON PC AND PC-LAN

W. J. KELLER AND J. G. BETHLEHEM,

Netherlands Central Bureau of Statistics

COMPUTER ASSISTED SURVEY PROCESSING ON PC AND PC-LAN

Wouter J. Keller and Jelke G. Bethlehem

Netherlands Central Bureau of Statistics, Automation Department
P.O. Box 959, 2270 AZ, Netherlands

Abstract

National statistical offices make an increasing use of computers in statistical information processing. Because of the enormous quantities of data that have to be processed, the microcomputer does not seem to be to most suitable instrument for this work. This paper shows that the contrary is true: use of microcomputers can increase the efficiency of the process and the quality of the produced information. Introduction of microcomputers has, however, an impact on organization of activities of both the automation department and the subject-matter departments. This calls for a strong policy with respect to decentralization of regular data processing, and standardization of automation tools. Software and hardware aspects are discussed, as is the Blaise system for computer assisted survey processing.

1. The statistical production process

The role of the computer in statistical information processing has always been an important one. Already in 1880 Herman Hollerith constructed his famous punched card machine, which was successfully used in the U.S. censuses of 1880 and 1890. And also for the U.S. Census Bureau one of the first electronic computers was built. In the beginning the computer was used for sorting, counting and tabulation only, but in the seventies a new generation of computers emerged, and statisticians obtained the possibility to carry out statistical analyses with packages like SPSS, SAS, P-STAT, and BMDP. Now the computer is also increasingly used in other parts of the production process: data editing, Computer Assisted Telephone Interviewing (CATI), and Computer Assisted Personal Interviewing with handheld computers (CAPI).

The production of statistical information is a complex process, in which data on persons, households and businesses, collected by means of surveys, have to be transformed into accurate statistics. The first step in this process is *data collection*, usually by means of a questionnaire. This ensures that all respondents are presented with the same questions. Data collected in this way will always contain errors. Therefore, in the second step a *data editing* process is necessary. The data are checked for errors, and detected errors are corrected as much as possible. After editing, the data are still not ready for analysis and tabulation. Nearly always the data are not representative for the population from which they originate. To repair this, a *weighting* procedure is carried out, which assigns adjustment weights to the sampled elements. Finally, a clean and representative data set is obtained, and then *tabulation* and *analysis* form the final step in the production process. This paper discusses the role of the computer in data collection, data editing, tabulation, and analysis.

Application of recent developments in computer technology is a prerequisite for efficient production of high quality statistics. At the Netherlands Central Bureau of Statistics (CBS) microcomputers are used increasingly in all steps of the statistical production process. Supplying a large organization with large quantities of microcomputers opens, on the one hand, new ways towards efficient statistical information processing. However, there are also problems which have to be dealt with. The CBS faces the challenge of managing decentralized information processing on hundreds of microcomputers. With respect to software, this calls for a strong policy on standardization of automation tools. In this paper, particular attention is

paid to the Blaise System for computer assisted survey processing, a tool which controls various steps in the statistical production process. But there also other problems which have to be solved: security, back-up, archiving, software licenses and sharing of programs and data files. Local area networks, and wide area networks are proposed as a solution.

2. Data collection

Statistical agencies collect most of their data by means of (sample) surveys. A questionnaire is defined, containing the questions to be asked to respondents. Traditionally the questionnaires are completed in face-to-face interviews: interviewers visit the respondents, ask the questions, and fill in the answers on the (paper) questionnaire. The quality of the collected data tends to be good. However, face-to-face interviewing is expensive. It requires a large number of interviewers, whom all have to do a lot of travelling. Therefore telephone interviewing has become popular. The interviewer calls the respondents from a statistical office, and no more travelling is necessary. Still, telephone interviewing is not always feasible: only people having a telephone can be contacted, and the questionnaire may not be too long and too complicated. A mail survey is still cheaper: no interviewers necessary at all. Questionnaires are mailed to potential respondents with the request to return completed forms. Although reminders can be send, the persuasive power of the interviewer is lacking, and therefore response tends to be lower in this type of survey.

During the last decade the computer is used increasingly in the data collection stage. First this occurred in telephone interviewing, and thus CATI-interviewing was born. The computer schedules the calls to make and controls the interview. It determines which questions are to be asked and checks the answers for consistency. This computer assistance has three major advantages. In the first place, it releases the interviewer from the burden of selecting the next question based on the answer of the current one and possibly present skip instructions. In the second place, this technology makes it possible to carry out checks during the interview. Detected errors can be discussed with the respondent, and hence corrected. This is not possible if error correction must be carried out afterwards, at the office. In the third place, the data are already entered into the computer during the interview, thus making a separate data entry phase superfluous. Nowadays CATI-interviewing is a well-established data collection technique. See for a full discussion of CATI-programs e.g. Nichols and Groves (1986[a], 1986[b]).

More recently, the advent of the small laptop computers made it possible for interviewers to take the computer along with them to the homes of the respondents. CAPI-interviewing combines the advantages of face-to-face interviewing and computer assisted interviewing. Experiments carried out by the CBS with the predecessor of the Blaise System (see Bemelmans-Spork and Sikkel, 1985[a] and 1985[b]), showed that interviewers were able to work with laptops, and that its use in face-to-face interviews did not cause any anxiety. Due to the success of these experiments, the CBS started full scale use of CAPI in a regular survey in 1987, i.e. the Labour Force Survey (see Van Bastelaer et. al, 1987[a]). Each month about 300 interviewers, equipped with laptops, visited 10,000 addresses. After a day of interviewing, the batteries of the hand-held computer are recharged at night, and the information collected that day is transmitted by telephone to the central office. The next morning, the interviewer finds a recharged machine with a clean workspace, ready for new interviews.

3. Data editing

Currently, the major data part is collected in a traditional way, on paper forms, and without the use of a computer. Completed questionnaires have to undergo extensive treatment. For producing high quality statistics, it is vital to remove errors. Three types of errors are distinguished: A *range error* occurs if a given answer is outside the valid set of answers, e.g. an age of 348 years. A *consistency error* is caused by an inconsistency in the answers to a set of questions. An age of 11 years may be valid, a marital status 'married' is also not uncommon,

but if both answers are given by the same person, there definitely is something wrong. The third type of error is the *routing error*. If the interviewer, or the respondent fails to follow the specified branch or skip instructions, i.e. the route through the questionnaire is incorrect: irrelevant questions are answered, and relevant questions are left unanswered.

Detected errors have to be corrected, but this can be very difficult if it has to be done afterwards, at the office. Particularly in household surveys, the respondent cannot be contacted anymore, so other ways have to be found to do something about the problem. Sometimes it is possible to determine a reasonable approximation of a correct value by means of an imputation technique, but more often an incorrect value is replaced by the special code indicating the value is 'unknown'.

Besides data editing, sometimes also another activity is carried out during this stage of the production process: *coding of open answers*. A typical example is the question which asks about the occupation of the respondent. Questions are most easy to process if a respondent selects one possibility from a list of pre-coded answers. However, for a question like occupation, this set of pre-coded answers will be very long, and it will be very hard for the respondent to select the proper answer. This problem is avoided by letting the respondent formulate his own answer, and copy the answer literally on the form. To be able to analyse this type of information, the answers must be classified afterwards, and this is a time consuming job, to be carried out by experienced subject-matter specialists. Presently, ways are explored to incorporate the computer in these activities. For the Labour Force Survey a computer assisted coding system was developed, see Van Bastelaer et.al. (1987[b]). This system is indeed called 'computer assisted' and not 'automatic'. The main reason for this is that a lot of the coding is not context-free, e.g. the code assigned to a particular occupation also depends on the description of the level of education and the height of the income. Furthermore, automatic text recognition techniques are not 100% successful.

4. Weighting

After data editing, and possibly coding of open answers, the result is a 'clean' file, i.e. a file without errors. However, this file is not yet ready for further analysis. In the first place, the sample is sometimes selected with unequal probabilities, i.e. businesses are selected with probabilities proportional to their size. A clever choice of selection probabilities makes it possible to produce more accurate estimates of population parameters, but only in combination with an estimation procedure which corrects for this inequality. In the second place, representativity may be affected by nonresponse, i.e. for some elements in the sample the required information is not obtained. If nonrespondents behave differently with respect to the population characteristics to be investigated, the results will be biased.

In order to correct for unequal selection probabilities and nonresponse, a weighting procedure is often carried out. Every record is assigned some weight, and these weights are computed in such a way, that the weighted distribution in the sample of characteristics like sex, age, marital status, and area reflects the known distribution of these characteristics in the population. Research has been carried out at the CBS in order to improve weighting techniques, see Bethlehem and Keller (1987). The increased power of the computer makes it possible to really implement these new techniques, see Bethlehem (1987).

5. Tabulation and analysis

Finally, we have a clean file which is ready for analysis. The first step in the analysis phase will nearly always be tabulation of the basis characteristics. A table looks like a simple thing, but in daily practice there is more to it. The composition of rows and columns (often built from more variables), the quantities displayed in cells (counts, means, percentages), the way in which percentages are computed, treatment of multiple response variables, the position of

totals and subtotals, and many other things, can make life very difficult. Good tabulation software packages can help a lot, but the usefulness of packages depends on their possibilities, the user-friendliness of the control language, and the possibility to produce camera-ready output.

Many statistical agencies also carry out analyses on their data, in order to reveal underlying structures, and thus gain insight in the data. Information obtained in this way, may improve a next survey, and thus improve quality or reduce costs. In many cases the data are categorical, and so many well-known statistical analysis techniques for interval data cannot be used. Of course, there are techniques for categorical data, e.g. loglinear analysis. Still, there is a need for other techniques which can also be applied on very large data sets. This has lead to the development and application of new techniques, like ANOTA or Correspondence Analysis. ANOTA (see Keller et.al. (1987)) is a simple alternative to logit analysis investigating the relationship between a dependent categorical variable and a set of explanatory variables. Correspondence Analysis, see Bethlehem and Sikkel (1987), is a technique to explore relations between categorical variables in a graphical way. Both techniques use bivariate tables as input instead of the raw data matrix, and therefore analysis on a large data set can even be carried out on a microcomputer.

6. The Blaise system

Blaise is the name of a new system for survey data processing, running on microcomputers (or a network of microcomputers) under MS-DOS. Blaise controls a large part of the survey process: design of the questionnaire, data collection, data editing and data analysis. The basis of the Blaise system is a powerful, structured language (also called Blaise) which describes the questionnaire: questions, possible answers, routing, range checks and consistency checks. With the thus specified questionnaire as input, the system automatically generates software modules for data processing. In the first place, a CADI-program (Computer Assisted Data Input) can be produced. This program provides an intelligent and interactive environment for data entry and data editing, of data collected by means of questionnaires on paper forms. Another important module is the CATI-program. The Blaise System automatically produces the software needed for telephone interviewing, including call management. A module which resembles the CATI-program, is the CAPI-program. This program for face-to-face interviewing with hand-held computers can also be generated by the Blaise System. Furthermore, Blaise automatically generates setups and system files for the statistical package SPSS. Interfaces to other statistical packages (e.g. PARADOX and STATA) are scheduled for the near future.

Traditional questionnaires are controlled by goto's, but this jumping makes the questionnaire unreadable. The Blaise language supports structured questionnaire design. This makes the structure of the questionnaire very clear. In almost one glance one is able to distinguish in which cases which questions are to be asked. Figure 1 gives an example of a simple questionnaire in Blaise.

A Blaise questionnaire is like a recipe for an apple pie. Such a recipe contains three parts:

Part 1: Preparing the ingredients
For the preparation of an apple pie first all ingredients must be put ready to hand. Likewise, in a questionnaire first all questions must be specified. This takes place in the *quest-paragraph*. A question consists of an identifying name, the text of the question and a specification of valid answers. Three types of answers can be distinguished in the example. Marstat and Job have user defined answer categories. Note that such answers consist of a short name, used for checking and routing purposes in other parts of the questionnaire, and an extended description, used for explication purposes in generated software modules. The question Age expects an integer answer which is restricted to the range 0, 1, 2, ..., 120. The question KindJob expects an open answer, i.e. any text of 80 characters maximum.

Figure 1. A simple Blaise questionnaire

```
QUESTIONNAIRE Demo;
QUEST
   Seq (KEY) "Sequence number of interview": 1..10000;
   Age       "What is the age of the respondent": 0..120;
   MarStat   "What is the marital status of the respondent":
              NotMar  "Never has been married",
              Married "Married",
              Divorce "Divorced"'
              Widowed "Widowed");
   Job       "Does the respondent have a job?":
             (Yes     "Yes, has a job",
              No      "No, does not have a job");
   KindJob  "What kind of job does respondent have?": STRING[80];

ROUTE
   Seq; Age; MarStat;
   IF Age > 15 THEN
      Job;
      IF Job = Yes THEN
         KindJob;
      ENDIF;
   ENDIF;

CHECK
   IF Age < 15 THEN MarStat = NotMar ENDIF;

ENDQUESTIONNAIRE.
```

Part 2: Mixing the ingredients
In the preparation phase all ingredients are mixed in the proper order. The analogue for the Blaise questionnaire is the *route-paragraph*. It describes in which situations which questions must be answered and the order in which the questions must be answered. There is a simple rule: writing down the name of a questions means asking it. In the example it is obvious that the questions Age and MarStat are asked to every respondent. Only persons older than 15 are asked whether they have a job. The question KindJob is only asked if the respondent is older than 15 and has a job.

Part 3: Tasting the result
After finishing the pie, it is tried out to establish whether the result is tasteful. Likewise, a questionnaire should contain consistency checks. In a Blaise questionnaire checks are specified in the *check-paragraph*. In particular, consistency checks are specified (range checks are automatically generated from the quest-paragraph and route checks are implied from the route-paragraph). The interpretation of this check-paragraph is: someone under 15 may not be married.

The description above does not exhaust the power of the Blaise language. In this general discussion, we have ignored the possibility of the block concept to build large and complex questionnaires from simple sub-questionnaires. We also did not mention the treatment of tables of questions in a spreadsheet-like way, local variables, user-defined answer-types, the use of external files, and variable texts, to name just a few. Details on these matters can be found in the Blaise reference manual (Bethlehem et. al, 1987[b]).

The Blaise System will also contain a module for interactive coding, which should be capable of handling the major part of the cases, leaving only a small number of hard cases, to be dealt with by a specialist. The module contains two different tools. The first tool implements a hierarchical approach to coding. Coding of an answer starts by entering the first digit of the code by selecting the proper category from a menu. After entering a digit, a subsequent menu is presented containing a refinement of the previously selected category. So, the description becomes more and more detailed until the final digit is obtained. The second tool contains a dictionary approach to coding. It tries to locate an entered description in an alphabetically ordered list. If the description is not found, the list is displayed, starting at the point as close

as possible to the entered description. The list can be made so that almost any description, including permutations, is present. The advantage of this method is that it is simple, fast and controlable. Both coding tools can be used simultaneously.

An important data processing application which can be produced by the Blaise System, is the CADI-machine. The CADI machine is an intelligent and interactive system for data input and data editing. Data can be processed in two ways. In the first approach the subject-matter specialist sits behind a microcomputer with a pile of forms and processes the forms one by one. He enters the data at the proper places and after completing the form he activates the check option, to test routing and consistency (range errors are checked during data input). Detected errors are highlighted and explained on the screen. Errors can be corrected, by consulting the form or calling the supplier of the information. After elimination of all errors, a clean record is written to file. If the specialist does not succeed in producing a clean record he can write the record to a separate file of problem records. These hard cases can be dealt with later on, also with the CADI-machine. In the second approach the data have already been entered beforehand by data typists, without error checking. With the CADI machine the records of the file can be edited. One by one a record is read from file, the check option is activated and, detected errors can be inspected and corrected.

The CADI-machine concentrates on processing data which have been collected on paper forms previously. However, the Blaise System can also be used for data collection. The system is able to produce the software, required to carry out CATI-interviewing or CAPI-interviewing. A CATI-program applies *dynamic routing*, i.e. the software takes control of the routing. After an answer has been entered, the CATI machine decides which question is to be asked next, and automatically jumps to that question. Note the difference with the static routing as supported by the CADI-machine. This difference arises because of the completely different use of CADI and CATI. With CADI, we want to copy a form and check what has gone wrong; with CATI, we want to avoid that anything goes wrong.

The CATI-machine applies dynamic error checking. As soon as a consistency error is encountered, an error message is displayed on the screen, together with a list of all questions involved. Via a menu, the user may select the question to jump to in order to correct the error.

The Blaise System can also produce a CAPI-program, an interviewing program to be used on a hand-held computer. CAPI combines the advantages of face-to-face interviewing with those of CATI interviewing. The program in the hand-held computer is in control of the interview, i.e. it determines the routing and checks consistency of answers. Since checking takes place during the interview, errors can be corrected on the spot. After successful completion of the interview, a clean record is stored in the hand-held computer. Later on, data is transmitted to the office by a modem and telephone, or a diskette can be send to the office.

7. Managing a 1000 microcomputers

A microcomputer is a valuable instrument for statistical data processing, but an efficient use requires a large number to be distributed throughout the organization. This is not without problems if stand-alone microcomputers are used. The first problem is security. Statistical information is based on individual data, and generally, individual data should be kept confidential. This can hardly be guaranteed if the data are stored on hard disks of microcomputers. A second problem is faced in backing-up and archiving the data, stored on diskettes and hard disks throughout a large building. Furthermore, there may be problems with software licenses: many microcomputers require many packages, and hence, many licenses. Also version control is not very easy. Finally, sharing programs, data files, and hardware is not possible in an efficient way.

The solution of the CBS of these problems is networking. Every deparment has its own local area network (LAN), running onder Novell software, and ARC-net hardware. Ten to fourty microcomputers are connected to a file server with a storage capacity of 150 to 300 Mb. Security is guaranteed by means of password protection in a login-procedure, by encryption, and by a write-lock for diskettes. In the future, diskettes will even become superfluous, and diskette stations can be removed completely. Since all software and data files are stored on the file server, sharing is now very simple.

The CBS has approximately 50 LAN's. For each LAN, a full-time LAN administrator and a deputy administrator are appointed. These two people are essential for installation and maintenance of software, and for user support. All LAN's are connected by a 'backbone' ARC-net (which will be replaced by Ethernet in the near future). Also the large CDC Cyber mainframe and the CDC 930 minicomputers are connected to this WAN, using Ethernet and bridges. The two sites of the CBS (Voorburg and Heerlen, which are 300 km apart) are connected by a 2 Mbps link.

Archiving and backing-up the LAN's is carried out in a centralized way by the Automation Department (total backup nearly 10 Gigabyte!). It is clear that version control and updating software can be realized very simple in such an environment. All software licenses are based on concurrent usage, which is controled by home-made software.

Allocation of microcomputers to departments is carried out by the board of directors, each half year, in batches of 200 microcomputers. Number of software licenses (in terms of concurrent 'counts') and documentation sets, amount of disk space, and number of printers are determined by simple rules based on the number of microcomputers in use allocated to the department. In order to cope with the problem of educating large numbers of new PC-users, the CBS runs an average of 50 one-day courses per month (occupying three fully equipped lecture rooms every working day).

8. Conclusions

The advent of the microcomputer has a considerable impact on the work of national statistical offices. More and more the unexperienced subject matter statistician is making use of it, and for his work he needs software like the Blaise System. On the one hand the power of this system lies in the consistency it enforces in the various steps of data collection and data processing. This makes the whole process more easy to manage and to control. On the other hand it also promotes standardization between different departments. Since all departments use the same software for carrying out their surveys, exchange of information between departments is more easy and less error prone.

Presently, the computer is fully integrated in the statistical production process. The CBS is equipped with hundreds of microcomputers. Each department has its own local area network (LAN), and all LAN's are connected with each other and with the mainframe and the mini's in the Automation Department. Simple and straightforward electronic data processing is now carried out by the subject matter departments themselves, leaving design and maintenance of complex information systems to be carried out by the computer specialists of the Automation Department. All interactive work (including data entry and data editing) is carried on microcomputers, while most batch-oriented work (weighting and tabulation) is concentrated on the mainframe and minicomputers.

The intensive and more sophisticated use of computers increases efficiency and data quality on the one hand, but on the other hand it also creates new problems. Due to the decentralization of computer activities there is a need for standardization of automation tools (only one wordprocessor, one database package, one spreadsheat, one analysis package, etc). In such an environment, training and support is vital, as is coordination. In other words: decentralized use in centralized environment.

References

Van Bastelaer, A.M.L., F.A.M. Kerssemakers and D. Sikkel (1987[a]): A test of the Continuous Labour Force Survey with hand-held computers: Interviewer behaviour and data quality. In: CBS select 4, Automation in Survey Processing (Netherlands Central Bureau of Statistics, Voorburg).

Van Bastelaer, A.M.L., L.M.P.B. Hofman and J.K. Jonker (1987[b]): Computer Assisted Coding of Occupation. In: CBS select 4, Automation in Survey Processing (Netherlands Central Bureau of Statistics, Voorburg).

Bemelmans-Spork, E.J. and D. Sikkel (1985[a]): Observation of prices with hand-held computers, Statistical Journal of the United Nations Economic Commission for Europe, vol. 3, no.2.

Bemelmans-Spork, E.J. and D. Sikkel (1985[b]): Data Collection with hand-held computers, Proceedings of the 45th session, International Statistical Institute, Book III, topic 18.3.

Bethlehem, J.G. (1985): LINWEIGHT User Manual. CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G., D. Denteneer, A.J. Hundepool, and W.J. Keller (1987[a]): BLAISE 1.1 / A first aquaintance, CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G., D. Denteneer, A.J. Hundepool, and W.J. Keller (1987[b]): BLAISE 1.1 / Language reference manual, CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G., D. Denteneer, A.J. Hundepool, and W.J. Keller (1987[c]): BLAISE 1.1 / The data entry machine, CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G., D. Denteneer, A.J. Hundepool, W.J. Keller, and M.H. Schuerhoff (1987[d]): BLAISE 1.1 / System description. CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G. and W.J. Keller (1987): Linear Weighting of Sample Survey Data. Journal of Official Statistics 3, pp. 141-154.

Bethlehem, J.G. and D. Sikkel (1985): CORAN User Manual, CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Keller, W.J., A. Verbeek and J.G. Bethlehem (1985): ANOTA: Analysis of Tables. CBS report, Netherlands Central Bureau of Statistics, Voorburg.

Nicholls II, W.L, and R.M. Groves, (1986[a]), The status of computer assisted telephone interviewing: Part I- Introduction and impact and cost and timeliness of survey data. Journal of Official Statistics 2, pp 93-115.

Nicholls II, W.L, and R.M. Groves, (1986[b]), The status of computer assisted telephone interviewing: Part II- Data quality issues. Journal of Official Statistics 2, pp 117-134.

THE USE OF MICROCOMPUTERS FOR CENSUS PROCESSING IN DEVELOPING COUNTRIES: AN UPDATE

V. TORO AND K. CHAMBERLAIN,

US Bureau of the Census

Introduction

This paper is an update of the paper entitled "The Use of Microcomputers for Census Processing in Developing Countries" written by Vivian Toro and Thomas Melaney and presented at the American Statistical Association meetings in August 1987. In it the authors discuss how microcomputers offer potential solutions to many of the problems developing countries encounter when processing census data with mainframe computers. The authors also describe the Integrated Microcomputer Processing System (IMPS) developed by the International Statistical Programs Center (ISPC), U.S. Bureau of the Census, as well as present three case studies of the use of IMPS by developing-country statistical offices.

The positive experiences of a number of developing countries confirm that microcomputers are technologically sound and cost-effective tools for processing censuses. Continuing technological advancements and refinement of microcomputer software make their use even more advantageous.

This paper focuses on the recent experiences of Burkina Faso and Senegal in using microcomputers. A summary is given of the use of IMPS by an increasing number of countries. Additionally, the paper describes the latest enhancements and future plans for IMPS. It concludes with some thoughts on trends on the use of microcomputers in the areas of data collection and processing, as well as in the use and dissemination of census data.

Microcomputers as Tools for Processing Census Data

Technological developments over the last 2 years have made the processing of large volumes of data on microcomputers even more practical. Problems normally associated with large-scale processing jobs such as population censuses are quickly disappearing. Data storage devices for microcomputers are more plentiful, faster, more reliable, smaller in size, higher in capacity, and less expensive. Processing speeds of microcomputers continue to increase and memory continues to expand.

Developing-country statistical offices have found this technology to be more accessible to them. For the most part, finding a local microcomputer vendor is no longer a problem. Typically, three or more different vendors are likely to be represented. This more competitive market has not only expanded the number of options, but also has reduced the prices and improved the terms of maintenance contracts.

Microcomputers are now an integral part of the operation of most developing-country statistical offices. They are being used for processing surveys, scheduling, budgeting, word processing, and other activities. A cadre of microcomputer literates exists in most of these countries. In addition, a variety of microcomputer software packages in different languages is now available.

Instead of asking whether or not to use microcomputers, statistical offices are now wondering how best to use them. They are asking what type of microcomputer software and hardware will best meet their short- and long-term needs. They are interested in obtaining the resources to buy this hardware and software, to train their staff, to provide hardware maintenance, and to integrate microcomputers to existing mainframe environments.[1/]

As statistical offices have become more comfortable with the microcomputer technology, many have chosen to use microcomputers for processing all or part of their census data. Burkina Faso, Senegal, and Micronesia were some of the pioneers that opted to process their census data using microcomputers. They were followed by countries such as Niger, Swaziland, Yemen, Honduras, Uruguay, and the Central African Republic. Countries including the Philippines, Pakistan, and Côte d'Ivoire chose to use microcomputers just for the entry of the census questionnaires data. Countries such as Tanzania and Cameroon chose to share processing between microcomputers and a mainframe computer.

Microcomputers have provided a solution to some of the problems encountered in the past by countries which used only mainframe computers to process their census data. Such countries have been able to afford the computing capability of the microcomputer within their limited budgets. The availability of several microcomputers has resolved most of the computer accessibility problems. Microcomputer maintenance is easier, and environmental standards are less restrictive. Additionally, training requirements are less demanding given the availability of software that is easier to learn and use, even for persons with little background in data processing.

The availability of microcomputer software which addresses the particular needs of census data processing has contributed greatly to the positive impact microcomputers have had in statistical offices. Most of the countries mentioned above have used or plan to use IMPS.

The Integrated Microcomputer Processing System: Background and New Features

BACKGROUND

The introduction of more powerful microcomputers in the early 1980's was not accompanied by microcomputer software which adequately addressed the requirements of census processing. ISPC embarked on the development of IMPS to provide microcomputer software for the major tasks of census data processing: data entry, editing, tabulation, analysis, and operational control. Because a project of this magnitude takes considerable time and resources, ISPC decided to use existing software where possible, and to provide intermediate products throughout the development of IMPS.

The design objectives of IMPS included ease of use, availability of a common data dictionary, ability to run with a standard microcomputer configuration, and modularity. The concept of modularity was of

particular importance in the early development of IMPS. The system was designed in such a way that modules could be easily swapped as software improvements were made. For example, prior to the development of the IMPS data entry component, CENTRY (Census Data ENTRY), commercial data entry packages were commonly used. The substitution of CENTRY does not require modification of the other application modules such as editing and tabulation. Another advantage of modularity is flexibility. The user can choose which tasks to perform using IMPS and which to perform using other software and/or other hardware.

The first step in the development of IMPS was to make the software packages CONCOR and CENTS 4 available on microcomputers. CONCOR (CONsistency and CORrection) and CENTS 4 (CENsus Tabulation System version 4) are editing and tabulation packages, respectively, developed by ISPC. These packages run on most mainframe computers and were in use by over 100 developing-country statistical offices by 1985. Both packages were written in COBOL, to ease the job of adapting them to the number of different mainframe computers found in the developing world. The availability of Realia COBOL, a powerful and efficient COBOL compiler for microcomputers, facilitated the adaptation of CONCOR and CENTS 4 to the microcomputer environment.

The first microcomputer versions of CONCOR and CENTS 4 appeared in 1984. The user instruction sets for these versions were the same as for the mainframe versions. This meant that the same CONCOR or CENTS application programs could run on either the mainframe computer or on microcomputers at statistical offices where both types of hardware were available. A number of statistical offices, including Senegal, Cameroon, and Somalia, took advantage of this flexibility.

Data capture, often a major bottleneck in census processing, was the next hurdle. ISPC evaluated commercial data entry packages to identify the best software for the capture of census and survey data. The evaluation identified two packages, ENTRYPOINT from Datalex, Inc., and RODE/PC from DPX, Inc., as meeting the performance criteria. At the time of the evaluation, RODE/PC had a slight advantage over ENTRYPOINT in terms of cost and speed. Therefore, ISPC recommended RODE/PC for census data entry on microcomputers.

The widely used Computer Programs for Demographic Analysis (CPDA), written by the Bureau of the Census in the early 1970's, was made available for microcomputers by Westinghouse Public Applied Systems under the sponsorship of the U.S. Agency for International Development (USAID). The Microcomputer Programs for Demographic Analysis (MCPDA) is one of several demographic analysis packages available on microcomputers. Currently, the Bureau of the Census is developing a more comprehensive package called Model Spreadsheets for Demographic Analysis (MSDA). Additionally, it is preparing a manual on the analysis of census data. This manual brings together the most useful methods of demographic analysis and projection, including brief descriptions, discussions of the advantages and disadvantages of different methods, and other helpful hints for interpreting the results. The manual will also discuss available software for performing

each type of analysis. The MSDA will be available in about 1 year and the manual during the following year.

The statistical analysis package PC-CARP, developed by Iowa State University, is the statistical analysis component of IMPS. Over the past 2 years it has become widely used by developing-country statistical offices for the calculation of variances in surveys, including post enumeration surveys. Users of PC-CARP include Zambia, Peru, Egypt, Zimbabwe, Costa Rica, Haiti, and the Philippines. The user interface of PC-CARP facilitates its use by analysts with little or no data processing experience.

With microcomputer software available for data entry, editing, tabulation, demographic analysis and statistical analysis, and with high capacity storage devices such as Bernoulli Boxes [2/] being announced, it became feasible to recommend census processing on microcomputers for countries with populations of 10 million or less. The 1985 census of Burkina Faso was the first major census to be processed entirely on microcomputers. RODE/PC, CONCOR, CENTS 4, and MCPDA were used. Burkina Faso's experience is described in detailed in the Case Studies section.

Since 1985, ISPC has added new features to IMPS. These include a common data dictionary, a more economical data entry module, improvements to CONCOR and CENTS, a census management and reporting system, and a census resource planning software.

NEW FEATURES

Common Data Dictionary

CONCOR and CENTS 4 originally were developed as stand-alone packages for mainframe computers. During their development, little attention was paid to integrating the two packages because the major concerns were portability and use of the software on computers with very limited memory capacities. As a result, the user was forced to define the data file for the CONCOR application program, then again for the CENTS 4 application program, using a different notation each time. After CONCOR and CENTS 4 were adapted to the microcomputer, the data definition became even more cumbersome because commercial data entry packages such as RODE/PC required yet a third definition of the data. Multiple data definitions were not only time-consuming to code but also error-prone.

The task of integrating CONCOR and CENTS through a common data dictionary resulted in CONCOR version 3 and CENTS version 5 which were released in early 1988. With these versions, the definition of the data file to be edited and tabulated is done only once. The interactive module of IMPS called DATADICT prompts the user for a name and the characteristics of each data item. The resulting definition, the Data Dictionary, is used by both CONCOR and CENTS.

In these new versions of CONCOR and CENTS, the user interface has been greatly simplified, and the CONCOR edit reports were made more concise. Several new features were added to CONCOR such as the ability to reference

repeating data items and greater flexibility in creation of the extract files.

CENTRY: Software for Data Entry

As the cost of commercial data entry packages rose, it became apparent that developing-country statistical offices needed a less expensive alternative. For example, in order to enter questionnaires within 1 year from a country with a population of 10 million would require approximately 30 microcomputers, meaning 30 copies of the data entry software package. At about $600 (U.S.) per copy, data entry software becomes quite expensive. Moreover, RODE/PC and ENTRYPOINT contain features, such as extensive logic checks, that are essential for survey data entry but not needed for census data entry. For census data entry, speed is of primary importance since the volume of data is so great. Capturing what is on the questionnaire, consistent or inconsistent, is essential. Inconsistencies can be corrected later through CONCOR and should not be left up to data entry staff to correct. Most developing countries have trouble just attaining acceptable keying rates with systems that have no or very limited logic checks.

Keeping in mind the particular needs for the entry of census data, ISPC developed a package called CENTRY as the data entry module of IMPS. CENTRY is a screen-oriented menu-driven package which allows for developing data entry applications, entering and verifying data, and collecting statistics on data entry operations. CENTRY uses the same Data Dictionary as CONCOR and CENTS. Features of CENTRY which make it attractive for census data entry include: programmable data entry screens, valid value checking, automatic duplication of fields, cursor control, skip pattern control, record retrieval and modification, and operator statistics. The data entered through CENTRY are stored in an ASCII file ready for use by other packages such as CONCOR. The user-friendliness of the IMPS data dictionary and CENTRY allows a person with minimal training who is familiar with the census questionnaire to set up a data entry application in a day. CENTRY requires 256 kilobytes of memory and two floppy disk drives. It is written in the C language to promote fast execution.

CENTRY will be released in November 1988. Already several countries, including Central African Republic, Burundi, and the Philippines, are planning to use it for census data entry.

dCONTROL: Census Management System

With a large data processing operation such as a national census, it is important to track each unit of data through the various processing phases. dCONTROL is an interactive census management and control system that helps census managers to monitor these phases. It also serves as a "check-in" facility for data by geographical unit, such as enumeration area (EA), thereby preventing duplication or omission of EAs. dCONTROL also allows for the production of preliminary count reports and other management reports.

In its current state, dCONTROL is a prototype that can be adapted to meet the particular needs of a country. It is being used in both Niger and Senegal where their statistical office data processing staff customized it to suit their specific needs. ISPC currently is developing a generalized version of dCONTROL which will be available in early 1989 in English, French, and Spanish.

dCONTROL requires a good cartographic system that accurately defines all the geographic and administrative boundaries for the census. The cartographic system serves as the basis for a geographic coding scheme that identifies each statistical area down to the EA level. dCONTROL monitors the various phases of processing at the EA level. The first phase is usually the receipt of the questionnaires from the field by the central office; the last phase is the keying of the census data.

dCONTROL is written in dBASE III PLUS programming language compiled using CLIPPER by Nantucket. It runs on an IBM PC/XT or PC/AT or compatible with 512 kilobytes of memory and at least 10 megabytes of auxiliary storage, usually a hard disk. The amount of hard disk needed depends upon the number of EAs in the country. For a country of 8,000 EAs, two megabytes of hard disk are sufficient to store the dCONTROL programs and the data.

CENPLAN: (CENsus PLANning)

Although great strides have been made in a very short time in the development of census processing software for microcomputers, there is still room for improvement. The goal is to make census results available as soon as possible after the census is taken. A number of factors contribute to the delay of census results, not the least of which is poor planning.

ISPC is developing a software package called CENPLAN which allows the census planner to determine the resources needed for census processing, given certain parameters such as population size, available computer equipment, and time constraints. Although CENPLAN cannot ensure that proper planning is done in time for a census, it provides the facilities to plan. CENPLAN uses spreadsheet software and will be available in French and Spanish, as well as English. The first release, which will address only the computer processing of a census, will be available in November 1988.

Case Studies

The number of countries using IMPS for the processing of census data continues to grow. Figure 1 is a partial list of countries that have used, are using, or plan to use IMPS to process their census data. Developing-country statistical offices are taking advantage of the system modularity that IMPS offers.

Some countries, including Niger, Senegal, Benin, Burkina Faso, the Central African Republic, Micronesia, Burundi, Yemen Arab Republic, Mali, Malawi, Swaziland, and the Pacific outlying areas of the U.S., have used or plan

to use microcomputers for all aspects of their census data processing. Senegal, Yemen Arab Republic, and Mali, had originally planned to use microcomputers for data entry and a mainframe computer for further processing of the data. However, when they realized they could obtain results in a more timely and cost-effective manner, they decided to process the data on microcomputers. For example, they found that CONCOR and CENTS generally execute faster on the microcomputers than on their mainframes. Even when the actual execution time was theoretically the same or slightly faster on the mainframe computer, they found they could obtain results faster on the dedicated microcomputers because they did not have to share computer resources with other users.

Other countries such as Somalia and Côte d'Ivoire have developed their CONCOR and CENTS programs using microcomputers but are running these programs against data on mainframe computers. Some countries like Ghana and Ethiopia did their data entry on microcomputers, then uploaded the data to a mainframe computer for processing. Furthermore, countries like Honduras, the Philippines, and Tanzania are still trying to decide whether to use microcomputers, a mainframe computer or a combination of the two for processing the census data. The selection criteria are not always solely technical. Some countries are committed to using mainframe computers for processing census data because their governments or donor agencies had purchased or procured the mainframe equipment several years ago with the express objective of using it for the census. Despite a technological revolution, amortization of this equipment was necessary before decision-makers could justify the purchase of new equipment.

IMPS modular design allows developing-country statistical offices to take advantage of existing hardware and software that is appropriate for census data entry. For example, Tanzania, and Cameroon are using minicomputers such as ICL/DRS 300 and IBM System/36 for data entry. On the other hand, Yemen Arab Republic is using microcomputers for data entry but has contracted out the development of their data entry programs because specialized Arabic data entry is needed. All these countries plan to use or are using CONCOR and CENTS for the editing and tabulation of the data.

Additionally, as other IMPS components become available, developing-country statistical offices can plan to take full advantage of it. For example, Burundi, the Central African Republic, and the Philippines are planning to use CENTRY for the entry of their census data.

Although designed primarily for census processing, IMPS also is being used for survey processing. The Gambia, Morocco, Zambia, Senegal, Portugal, Ghana, Rwanda, Egypt, and American Samoa are using one or more of IMPS modules for surveys.

The following two case studies, Burkina Faso and Senegal, were presented in the Toro/Melaney paper as representative of the positive impact of IMPS on census processing activities. An update of their experience follows. The Federated States of Micronesia (FSM), the third case study presented in the Toro/Melaney paper, have continued to use microcomputers successfully to process the census data for some of the states. FSM is

currently discussing the establishment of a national census that would standardize the date, questionnaire, and processing activities used by all the states.

Burkina Faso

In December 1985, the National Institute of Statistics and Demography (INSD) of Burkina Faso, formerly Upper Volta, conducted their second national census. It took them only 18 months to enter and process the data of their estimated population of 9,000,000.

Three IBM PC/AT microcomputers with 512 kilobytes of memory, and 20 megabytes of hard disk storage were used to process the data. The data entry equipment consisted of 22 IBM PC microcomputers each having 512 kilobytes of memory, and two IBM PC/XT microcomputers with 10 megabytes of hard disk storage and 640 kilobytes of memory. Related peripherals included six printers, two DIGIDATA tape drive units, and three Bernoulli Boxes with 10 megabyte removable cartridges. The tape drive units were used for backup and archival purposes, while the Bernoulli Boxes were used for primary storage of the census data. RODE/PC, CONCOR, and CENTS 4 were the software packages used for data entry, editing, and tabulation, respectively.

The INSD data processing staff for the census project consisted of two full-time programmers and a long-term United Nations advisor. In addition, the Bureau of the Census under an agreement with USAID, provided data processing assistance in the form of training, program development, and monitoring of the RODE/PC data entry application program, the CONCOR editing program, and the CENTS 4 tabulation programs.

The preliminary counts report based on the manual counts was available by March 1986. Two shifts of 25 operators began keying the 15-percent sample of the data in September 1986. Two months later, this data capture operation was complete, and by March 1987, the tabulations for the 15-percent sample had been produced. The edited sample data occupy 11 Bernoulli cartridges of 10 megabytes each. (It should be noted that data entry was begun several months late due to delayed arrival of microcomputer equipment.)

The entry of the remainder of the data took place between December 1986 and May 1987. The 40 tables for the entire country were available by March 1988, 18 months after data entry began. The keyed data for each enumeration area were stored on one diskette. No more than 600 diskettes were in use at one time since they were recycled after data verification. After the data were completely verified, all the EAs for a province (average 200 EAs per province) were transferred onto one or more Bernoulli cartridges. Approximately 10,000,000 records (9,000,000 population and 1,000,000 housing) of 52 characters in length each were saved. Sixty Bernoulli cartridges of 10 megabytes each were used to store one copy of the country's data. Five versions of the country data were kept.

Figure 1. Partial List of IMPS Users for Population Censuses

Country Name	Census Date	Est. Pop. (millions)	Operational Control	Data Entry	Editing	Tabulation
Benin	1989	4.3	manual	CENTRY	CONCOR	CENTS
Burkina Faso	12/85	9	part autom.	RODE/PC	CONCOR	CENTS
Burundi	8/90	5	undecided	CENTRY	CONCOR	CENTS
Cameroon	4/87	10.3	manual	IBM S/36	CONCOR	CENTS
Central African Republic	10/88	2.7	dCONTROL	CENTRY	CONCOR	CENTS
Comoros	1990	.5	manual	undecided	CONCOR	CENTS
Côte d'Ivoire	3/88	10	manual	RODE/PC	CONCOR/a	CENTS/a
Honduras	5/88	4.8	dCONTROL	KeyEntry III	CONCOR/b	CENTS/b
Malawi	9/87	7.4	manual	ICL mini	CONCOR	CENTS
Mali	4/87	8.4	manual	RODE/PC	CONCOR	CENTS
Mauritania	10/87	1.9	manual	UNKNOWN	CONCOR	CENTS
Micronesia (Pohnpei)	9/85	.30	manual	Entrypoint	CONCOR	CENTS
Niger	5/88	7.5	dCONTROL	RODE/PC	CONCOR	CENTS
Philippines	1990	57	undecided	CENTRY	CONCOR/b	CENTS/b
Senegal	5/88	7.5	dCONTROL	RODE/PC	CONCOR	CENTS
Somalia	11/86	7.8	manual	custom pgm	CONCOR/a	CENTS/a
Swaziland	8/86	.7	manual	RODE/PC	CONCOR	CENTS
Tanzania	8/88	23.5	manual	ICL mini	CONCOR/b	CENTS/b
U.S. (Pacific Outlying Areas)	4/90	.3	manual	CENTRY	CONCOR	CENTS
Yemen A.R.	2/86	9	part autom.	custom pgm	CONCOR	CENTS

a/ Program development done on microcomputers; production processing on mainframe computers.

b/ Program development done on microcomputers; have not decided what to use for production processing.

They include the raw data at enumeration area level (RODE/PC file format and ASCII format), the consolidated data at province level, and two copies of the edited data.

The management of about 600 diskettes and 240 Bernoulli cartridges was accomplished effectively by using a simple external labeling system. Each diskette was clearly identified by its external label indicating an enumeration area. Equally, every Bernoulli cartridge was identified by an external label indicating a province. The task of managing and processing the census data was manageable but inconvenient, particularly because the Bernoulli Boxes were capable of holding only 10 megabytes of data at a time, and the IBM PC/AT microcomputers had only 20 megabytes of hard disk storage. Due to technical difficulties, the tape units could only be used for archival purposes and not for processing the census data.

CONCOR and CENTS do not require the entire population census file to be processed or sorted as a unit. The data can be edited batch-by-batch using CONCOR. The batches can vary in size. The edited (or cleaned) batches of data can be merged into larger batches (or geographic units) for tabulation using CENTS. The resulting cross-tabulations can be saved as much smaller files, and then consolidated to produce tables for larger geographic entities. Thus, in a system using more than one microcomputer, each machine can be processing a different geographic area simultaneously, shortening the overall processing time.

In the case of Burkina Faso, both editing and tabulation were done at the province level. The CONCOR program took an average of 20 minutes to edit each of the 30 provinces. The execution of the CENTS tabulation programs for each province took about 15 minutes. Through CENTS, the intermediate province-level tabulations were merged to produce the country-level tabulations. Three IBM PC/AT microcomputers with 20 megabytes of storage were used for editing and tabulation.

The keying of the country data in 9 months instead of the estimated 1 year was due primarily to the high keying rates attained. The keystrokes ranged from 9,000 to 14,000 per hour with an error rate of .3 percent. Since INSD was not able to give monetary incentives to the keyers, they provided them with other types of motivation. For example, the best keyers could choose to work during the shift (morning or evening) that was more convenient to them and schedule their leave more freely. They also were promised a permanent job at the INSD after the census was over. Given the high unemployment rate in Burkina Faso, most keyers were motivated just to keep a relatively well paying job. Out of 500 applicants only 50 were selected after taking a written test and being interviewed. Above all, the keyers were hard working and took pride in their work.

Maintenance of the microcomputers was accomplished by the data processing team. A supply of spare parts allowed them to keep downtime to only 4 percent. The diskette units and the keyboards caused the most problems. In spite of these and some initial electrical problems, the equipment maintenance was done in a timely and nondisruptive fashion.

As a pioneer, INSD demonstrated that processing a population census of 9,000,000 persons using microcomputers is not only feasible and cost-effective, but an effective approach to the production of timely results.

Senegal

The Direction de la Statistique (DS) took the second national population census of the Republic of Senegal in May 1988. The estimated population was 7,500,000.

The DS chose to purchase microcomputers for the data entry operation instead of renting IBM 3742 equipment because of the reasonable cost of microcomputer hardware, the availability of adequate data entry software, and the potential use of microcomputers for other projects after the keying operations are finished. The editing and tabulation of the data were scheduled to be done using an IBM 370/145 at the Computer Center of the Ministry of Finance, a data production facility used by other government agencies.

The original microcomputer configuration included 18 IBM PC microcomputers with 256 kilobytes of memory for data entry, two IBM PC/XT microcomputers with 10 megabytes of hard disk storage for program development, five printers, and a tape unit used for data backup. Uninterrupted Power Supply (UPS) units are used for power supply protection.

The RODE/PC software package was selected for data entry. The data would be edited using CONCOR and tabulated using CENTS 4. The CONCOR and CENTS 4 programs could be developed using the microcomputers, then transferred to the mainframe computer for production runs.

During a 3-week technical assistance visit to Dakar in January 1986, data processing advisors from ISPC, under an agreement with USAID, conducted a CENTS 4 workshop, installed CONCOR and CENTS 4 on the IBM mainframe computer, and discussed tabulation plans and specifications with DS personnel.

Originally, the data processing staff responsible for the census operation consisted of the head of the data processing section and the head of the data entry section of the DS. They furthered their training during a work-study visit in early 1986 at ISPC in Washington, D.C. During a 6-week stay, the data entry supervisor, who had no programming experience, learned to use menu-driven RODE/PC and assisted in developing the data entry application for the census questionnaire. The data processing chief, an experienced systems analyst with prior CONCOR and CENTS 4 training, remained at ISPC for 3 months. In addition to learning RODE/PC, she wrote the CONCOR edit and imputation programs. She also completed much of the tabulation programming using CENTS 4. All the work was done on an IBM PC/XT.

Upon their return to Senegal, they were able to implement modifications to the programs easily when significant changes were made to portions of the questionnaire. The data for the pilot census, conducted in March 1987,

were keyed using microcomputers. The CONCOR and CENTS 4 programs were transferred to the mainframe computer. No changes to the code were required. The keyed pilot census data also were transferred from 5 1/4-inch diskettes to 8-inch diskettes to facilitate transfer to the mainframe computer.

Although they encountered no major problems while processing the pilot census data on the mainframe computer, the data processing staff became frustrated due to the inconvenience. They had to physically transport themselves and the data to the mainframe computer site several miles away. After conducting a test, they found that because of the need to share the computer facilities, it takes longer to run CONCOR and CENTS 4 programs on the mainframe computer than it does on the dedicated microcomputers. The data processing staff also found the interaction with the IBM PC much friendlier than with the IBM mainframe computer.

As a result of their growing positive experiences with microcomputers, the data processing staff decided to process the data for the full census using microcomputers instead of using mainframe computers. To upgrade the current microcomputer configuration to allow in-house processing for the full census, the data processing staff acquired extended mass storage consisting of three 20-megabyte Bernoulli Boxes and a tape drive unit. Two of these Bernoulli Boxes contain a hard disk with 80 megabytes of memory. Although the file management of diskettes, Bernoulli cartridges, and tapes will be a critical component of the processing, the data processing staff have judged the file management problems less formidable than the inconveniences of working in the mainframe computer environment. Additionally, the head of the data entry section feels confident about the handling of census data files because he developed and implemented the magnetic media management system used for the previous population census.

The DS acquired six additional IBM PC/AT microcomputers. Four are being used by the data processing staff for program development and for production processing. The other two microcomputers will be used by the demographers and statisticians for the analysis of the census data. DS plans to use the MCPDA for the demographic analysis of the census data.

The head of the data processing section has transferred her microcomputer expertise to three system analysts who also are working on the processing of the census data. Additionally, ISPC has been providing technical assistance to the DS in all aspects of the census processing activities. One of these areas is the monitoring of census processing activities and the production of management reports, including the preliminary counts report. The DS staff adapted the dCONTROL prototype to meet Senegal's specific needs. The preliminary counts reports which are based on manual counts of the May/June 1988 enumeration were available by October 1988. The data base and dCONTROL programs occupied less than two megabytes.

The DS plans to process a 10 percent sample of the data first. However, the availability of a permanent site for storing the questionnaires and for coding and keying the data has delayed the start of the processing of the census data until mid-September. The RODE/PC data entry program and

the CONCOR editing programs have been tested. The CENTS tabulation programs are being finalized. A system to monitor operator performance based on operator statistics such as keystrokes per hour and keying error rates has been developed. This dBASE III PLUS system produces reports with operator statistics based on the information found in the ASCII file produced by the RODE/PC data entry program.

In the meantime, the DS staff have been using RODE/PC and CENTS for other surveys. The DS staff's enthusiasm for microcomputer use extends even further. As part of the analysis and publication stage, the DS plans to use microcomputers for thematic mapping. This is the computerized creation of maps that reflect regional variations in characteristics of the population. It requires mapping software and peripheral equipment such as a digitizer and a plotter. The U.S. Bureau of the Census will provide training.

Over the last 3 years, the DS has become more confident and enthusiastic about microcomputers. Not only are they using microcomputers to process the census data, but they also are using them for many other projects and surveys. Most systems analysts have microcomputers on their desks and are providing microcomputer training to junior programmers. Demographers and statisticians also are using microcomputers for their work.

With the subject-matter expertise gained during the last census, the growing microcomputer literacy, and staff enthusiasm, the prognosis for the census processing is good.

General Prognosis

The experiences of Burkina Faso and Senegal are representative of the positive impact microcomputers and IMPS are having on the census processing activities of many countries. Three years ago these countries were pioneers in the use of microcomputers for censuses. Now over 18 countries are using or plan to use microcomputers for some aspect of their census processing. This number is increasing rapidly as statistical offices and donor agencies realize the advantages of microcomputers.

Operational control and file management are still a challenge. As recording densities of mass storage devices continue to increase, operational control and file management will be easier, perhaps even easier than on mainframe computers. The dCONTROL census management and reporting system is helping to resolve this challenge.

In spite of improvements, data capture is still the most time-consuming operation in census data processing. There is a limit to the rate at which data can be keyed, regardless of the speed of the software. Although one could add work stations to diminish the total time for keying, there is a point at which the logistics of such an operation become highly cumbersome. As optical mark reading (OMR) technology advances, it should become a more cost-effective data entry alternative for developing countries within the next 5 years.

The availability of more cost-effective tools for processing census data should make it possible for developing-country statistical offices to focus their attention and resources on the use and dissemination of census data. Many tools are becoming available to accomplish this objective, including software packages for the production of thematic maps and desktop publishing. Additionally, hardware improvements in optical disk technology allow for easier access and transfer of large volumes of data.

Census data users themselves have more tools available which should, in principle, help relieve statistical offices of some of the responsibility they have as the principal disseminators of census data. Since many census data users have never been able to obtain census results or have obtained them 5 or more years after enumeration, rendering their usefulness almost null, the availability of any data is a big improvement. Census data users will welcome the ability to access a subset of the census data on diskettes that they can use on their own microcomputer systems. Hopefully, statistical offices and census data users will focus on the analysis and dissemination of reliable data and not be distracted by the many presentation alternatives provided by the new technology.

As computer technology advances, data processing needs and software are changing. IMPS is the result of years of experience by ISPC staff, but is also the outcome of requests by census data processing managers in developing countries. As resources allow, ISPC will continue to support and modify the various modules of IMPS by critically observing its use by developing-country statistical offices, listening to IMPS users regarding suggestions for improvements, and by taking full advantage of the microcomputer environment available in developing country statistical offices.

ISPC's primary objective is to make these packages usable by persons who are not highly skilled or experienced computer programmers. ISPC will make modifications to CONCOR and CENTS in the area of user-friendliness over the next few years. One goal is to make at least the definitional aspects of these packages interactive so that users need not learn a procedural language in order to use them. Some aspects of the packages, such as the definition of complex consistency edits are best stated through a procedural language. Others, such as the definition of table formats, are best done interactively. ISPC also will develop tutorials and extensive census examples to facilitate the learning process.

Enthusiasm for new technology should not diminish the effect that planning, politics, and communication have on the processing of a census. After all, microcomputers are only tools for people to use; they are not a substitute for careful planning and close coordination.

Conclusions

The experiences of developing countries like Burkina Faso and Senegal have made the use of microcomputers to process censuses of 10,000,000 persons a reality. Technological developments over the last 3 years are making some of the problems normally associated with large-scale processing quickly disappear. As more cost-effective tools become available for processing census data, developing-country statistical offices should be able to focus their attention and resources on the use and dissemination of census data. We eagerly await the impact of future technological development on large-scale data processing.

NOTES

1/ In this paper, any reference to mainframe computers will also include minicomputers.

2/ Bernoulli Boxes are removable cartridge-based hard-disk systems. They can serve as both mass-storage devices and backup systems. Information is stored in greater densities and higher data access speeds are allowed. Each cartridge unit has one or two cartridge readers. Depending on the model, the units can read data from either 10 or 20 megabytes cartridges. Since these cartridges are removable, they provide virtually unlimited capacity. The newer models also have up to 80 megabytes of hard-disk storage in addition to the removable cartridges. In addition, they are physically very resistant to rough treatment which would otherwise have destroyed or caused the loss of data on hard disks. They have been used for several years in developing countries and have proven to be very reliable.

REFERENCES

Anderson, Patricia, and Ondra, Thomas. 1987. Trip and assessment report to Pohnpei, Federated States of Micronesia. U.S. Bureau of the Census.

Chamberlain, Kathleen. 1988. Trip reports to Mogadishu, Somalia. U.S. Bureau of the Census.

Chamberlain, Kathleen. 1988. Trip reports to Dar Es Salaam, Tanzania. U.S. Bureau of the Census.

Gomez, Enrique. 1988. Trip reports to Tegucigalpa, Honduras. U. S. Bureau of the Census.

Gomez, Enrique, et al. 1987. Trip reports to Tegucigalpa, Honduras. U.S. Bureau of the Census.

Hie, Jean-Marc. 1988. Trip report to Ouagadougou, Burkina Faso. Economic Commission for Africa. United Nations.

Le, Thanh, and Toro, Vivian. 1988. Trip report to Dakar, Senegal. U.S. Bureau of the Census.

Le, Thanh. 1988. Trip reports to Bujumbura, Burundi. U.S. Bureau of the Census.

Megill, David, and Toro, Vivian. 1988. Trip report to Niamey, Niger. U.S. Bureau of the Census.

Megill, David. 1988. Trip report to Niamey, Niger. U.S. Bureau of the Census.

Melaney, Thomas, and Toro, Vivian. 1988. Trip report to Dakar, Senegal. U.S. Bureau of the Census.

Rowland, Sandra, and Melaney, Thomas. 1988. Trip report to Bujumbura, Burundi. U.S. Bureau of the Census.

Sawaya, Selma, and Banerjee, Kakali. 1987. Trip report to Abidjan, Cote d'Ivoire. U.S. Bureau of the Census.

Stroot, Michael. 1988. Trip report to Sana'a, Yemen Arab Republic. U.S. Bureau of the Census.

Toro, Vivian, et al. 1987. Trip report to Niamey, Niger. U.S. Bureau of the Census.

Toro, Vivian, et al. 1987. Trip report to Dakar, Senegal. U.S. Bureau of the Census.

Toro, Vivian, and Melaney, Thomas. 1987. "The Use of Microcomputers for Census Processing in Developing Countries," paper presented at the American Statistical Association Meetings, San Francisco, California.

U.S. Bureau of the Census. 1988. "Population Census Support Status of Sub-Saharan Africa"

A REPORT ON RECENT SURVEY PROCESSING IN DEVELOPING COUNTRIES: THE DEMOGRAPHIC AND HEALTH SURVEYS MICROCOMPUTER APPROACH

J. CUSHING,

Institute For Resource Development/Westinghouse

1. Introduction

The Demographic and Health Surveys Program (DHS), being implemented by the Institute for Resource Development/Westinghouse (IRD), is funded by the U.S. Agency for International Development (US/AID). The first five-year phase of the DHS, begun in October 1984, has been providing financial and technical assistance for 35 surveys in 30 lesser-developed countries (LDCs). The second phase of the program, which includes 25 additional surveys, was recently awarded to IRD, allowing the program to continue for 5 additional years without interruption.

The surveys conducted under the DHS Program have had sample sizes ranging from 3,000 to over 12,000 individual respondents. The core questionnaires, of which there are two, include a household roster and an eight-section individual questionnaire applicable to eligible women listed in the household roster. The individual questionnaire collects data on background characteristics of the respondent and spouse (if applicable), fertility (including a full birth history), contraceptive knowledge and practice, fertility preferences, maternal and child health for births in the last 5 years, and the heights and weights of surviving children 3 to 36 months. The two core documents differ in that the "A-core" expands the contraceptive practice section while the "B-core" expands the health section. A flat file of individual respondent data requires approximately 2,000 characters per case.

The first DHS was conducted in El Salvador in early 1985. In order to have a data file for preliminary tabulations, a subset of variables was coded onto forms, scanned onto a microcomputer, written to a mainframe and tabulated. The full survey was processed in El Salvador by traditional mainframe methods using data transcription, blind data entry, and cyclical batch editing as shown in Figure 1.

Microcomputer processing had been considered for the DHS in El Salvador, but it was rejected because of its relatively untested capabilities when applied to a large-scale survey like the DHS. In addition, mircrocomputer software appropriate for processing survey data was still under review. By the fall of 1985, however, DHS committed itself to microcomputer processing, recognizing that its timetable would not permit the delays generally experienced in LDC mainframe processing. Commercially available software was used while the DHS Program developed an integrated processing system that could alleviate the processing bottlenecks experienced by earlier survey efforts in developing countries.

As a result of these efforts, the DHS now completes data entry approximately two weeks after the end of fieldwork. At that point, the data file is structurally correct, contains no range or skip errors, and has few inconsistencies remaining within questionnaires. Secondary batch editing of complex internal consistencies may continue for another month or two. Preliminary reports are now being produced in less than three months after the end of fieldwork. Delays in final reports are no longer due to data processing; rather they are caused by the time required to review, approve, and produce the report. Nevertheless, final reports are being published approximately one year after the end of fieldwork.

Table 1 presents the cumulative efforts during DHS, Phase 1. It should be noted that preliminary reports were not specified for 4 surveys; 2 surveys were experimental in nature with a much longer report preparation time; and 5 surveys are currently in the field. Figures in parentheses are estimated numbers for the remainder of 1988 and 1989.

Table 1. The Cumulative Number of DHS Surveys and Reports by Calendar Year

ACTIVITY	1985	1986	1987	9/1988	1988	1989
Surveys fielded	1	10	23	29	(34)	(35)
Preliminary Reports published	0	1	12	20	(22)	(31)
First Country Reports published	0	0	3	11	(21)	(35)

An examination of Table 1 shows that production of reports for the earliest surveys (through 1986) was considerably slower than current production. Preliminary reports for a few surveys fielded early in 1986 were not available until 1987. However, preliminary reports for surveys fielded in 1988 have been produced within the three-month period following fieldwork. Final reports for all surveys fielded in 1987 have been published or are projected for publication before the end of 1988.

On average, surveys have been in the field three to four months. Preliminary reports have been published three months later in Latin America, North Africa and Asia, and five months later in sub-Saharan Africa. Most First Country Reports have been published approximately one year after the end of fieldwork; a few have required 18 months to two years. This can be compared to earlier survey efforts which often required from two to four years for the publication of results.

The improvement in lag time from the end of fieldwork to published reports can be credited to several factors: the move from a large, multi-interest mainframe environment to a self-contained unit within the survey organization; the development of an integrated system for survey processing and analysis (ISSA); and a change in the philosophical approach to processing a survey.

2. Moving to a Self-contained Microcomputer Environment

The problems of a mainframe environment in many LDCs can be especially troubling to social science researchers. All users may suffer lost time due to downtime of obsolete hardware, a lack of appropriate software and/or an insufficient number of skilled programmers; researchers often have the additional problem of being relegated to the end of a long processing queue. Programmer and machine time can be difficult to gain because of the computer needs of government administrative and finance units.

Equipping a survey organization with two to three microcomputers, peripherals such as printers, Bernoulli Boxes and voltage regulators and user-friendly software removes the problem of computer availability from survey processing. An additional benefit is the upgrading of report generation capabilities through word processing and graphics software. The cost of providing the hardware, software and supplies can be less than the cost of mainframe processing in some countries.

DHS has experienced very few hardware problems during or after survey processing. In general, equipment has been purchased three to six months prior to anticipated fieldwork dates and used in the DHS home office during this testing period. Most hardware problems, such as Bernoulli Box, hard disk or keyboard failures, have occurred there and repair or replacement has been easy.

Hardware configurations for the early 1986 DHSs consisted of IBM/XTs or COMPAQ portables, NLQ printers and UPSs. As hardware costs dropped and processing speed increased, ATs, 286/12 Mhz

micros, and 386/20 Mhz micros have been provided. Today's configuration for each survey consists of one 386 unit and two 286 units, one Bernoulli Box, printers and UPSs or voltage regulators. All micros are equipped with 30 to 70 Mbyte hard disks.

Although the environmental concerns for micros are considerably less than those for a mainframe, DHS has required that the room designated for computer activities be clean and dry, since dust and humidity cause more diskette and hard disk problems than any other factors. Regulations regarding food, drink, smoking and cleaning in the computer room have helped to ensure that the equipment will continue to function well.

Survey-implementing agencies in some countries have had knowledgeable microcomputer programmers available to them. Nevertheless, DHS usually has included training of survey personnel in the maintenance and use of equipment and supplies. The training focus, however, has been in the use of software needed for processing the survey, such as DOS and data entry, editing and tabulation software. Portions of the DHS Data Processing Manual supplement the specific hardware and software user's manuals.

DHS has provided each implementing organization with SPSS/PC+ for data analysis, a word processing package, such as Volkswriter or Word Perfect, and utility software, such as Norton's Utilities or PC Tools. More recently, graphics packages have also been installed to upgrade dissemination capabilities. Unfortunately, sufficient time for thorough training in these packages has not often been available. User interest has been a strong impetus for self-learning.

Clearly, additional training is necessary when survey processing is moved to a microcomputer environment. DHS has installed its own menus to assist computer novices in accessing DOS, word processing, graphics, utility and data processing packages. To simplify processing, DHS elected to develop an integrated package capable of data entry, editing, imputation, tabulation and file construction.

3. Developing an Integrated System for Survey Analysis (ISSA)

The DHSs fielded in early 1986 used a combination of software packages: Entry Point for data entry, CONCOR for data editing and SPSS/PC for tabulations. As development of ISSA progressed, it was also used for some editing and tabulations in these early surveys. The use of various packages with different file format requirements and command languages reinforced the need for one integrated system capable of data entry, verification, editing, tabulation and file construction.

The intelligent data entry component of ISSA was designed to ensure that a user could produce a structurally correct data file with no range, skip or consistency errors within a case. This is possible because of the data structure design within ISSA, which keeps an entire case available for change during data entry and protects the data path within the case. No operator overrides of ranges or skips are permitted.

Without these controls, processing a survey with "intelligent data entry software" on a micro could revert to the same time-consuming round of secondary batch editing that traditional processing methods require. Structural, range, skip and consistency errors might exist after data entry. Editors would need to return over and over again to questionnaires, searching for corrections to be entered into the data file.

If different packages were used for data entry and editing, the issue would be complicated because of the need to write the same edits in two different programming languages. Although there

would be fewer errors to correct than in a heads-down-keying data entry environment, the editing would have to be done and the amount of time devoted to programming and editing could be considerable.

Entry Point and CONCOR, used by DHS for data entry and editing in 1986, provide an example of this kind of problem. Although Entry Point permitted range and skip edits during data entry, they could be overriden by the data entry operator. Range errors were flagged in the data file, but the Entry Point option to locate the flagged fields was not always used. Thus, out-of-range values and inconsistent data paths appeared in the data and required further editing with CONCOR.

Skips based on fields located on previous screens and data file structure could not be rigorously controlled because data entry operators could not return to previous screens to verify that correct values had been entered in fields that controlled the skips. For example, a field containing the number of live births for a respondent could not be used to control the number of entries in a birth history because it was not possible to return to the field if it had been erroneously keyed. It was necessary to include all structural edits in a secondary CONCOR edit program and return to the Entry Point data file for corrections.

Data entry programs that allow operator overrides often flag the fields that contain errors. Because of the need to store a flag with each field, data storage may be in a format specific to the package; before using the data with other packages, it will have to be transcribed into an ASCII data set. One of the most time-consuming activities during the processing of the 1986 surveys was the transcription from Entry Point format to ASCII for CONCOR processing. DHS rejected this format option for ISSA data files; the data files that are created by ISSA are in ASCII format.

Many data entry packages process data files sequentially; to retrieve the 500th case requires the reading of the first 499 cases. ISSA produces an accompanying index file for each data file created during data entry. If destroyed, the index can be recreated through an ISSA utility. The index allows immediate retrieval of a case for updating within ISSA.

An ISSA dictionary, which defines the data file for data entry activities, can be used with the batch components of the system: verification, editing, imputation, tabulation and file construction. The dictionary includes ranges of data values, variable labels and value labels. Its construction and use are completely menu-driven; it includes a form editor for designing data entry screens.

The tabulation component of ISSA was developed to allow production of all tables required for DHS reports. Simple crosstabs or means, composite crosstabs or means consisting of multiple row and/or column variables, and tables constructed from multiple table manipulations can be produced by ISSA. In order to speed batch processing, ISSA can transpose a data file of cases to a data file of variables. A benchmark test of several crosstabulations of a 6,000-case file was completed on a 386 micro in approximately one minute.

The DHS questionnaire produces a hierarchical data file of two levels: household and respondent. Since many researchers may want to access SPSS/PC+ for statistical analysis, ISSA provides a utility to produce a flat data file that can be read by SPSS/PC+. A file of SPSS/PC+ commands to read the file and label the variables and values is also produced.

In addition to the data collected at the household and individual levels, DHS has collected data at the primary sampling unit (PSU) level. To incorporate the community level data with the standard DHS household/respondent data file, ISSA was designed to process multiple files. Up to three external indexed files can be read and updated while processing a master input file.

The needs of DHS demanded the development of a strong data entry, editing and tabulation package. By the end of 1986, ISSA was meeting almost all of the data processing needs of DHS. Further development of ISSA is planned for Phase 2 of DHS; this will include more user-friendly tabulation capability and calculation of descriptive sampling errors.

4. A Change in the Philosophical Approach to Processing a Survey

Clearly, the availability of microcomputers and the development of powerful software have contributed strongly to the success of the DHS Program; but without a change in the philosophical approach to survey processing, the impact would not have been as great. Both the micros and the software could have been used to continue processing in the traditional manner shown in Figure 1, depending upon secondary batch editing for production of a clean data file. DHS, however, chose to adopt a different approach as shown in Figure 2, emphasizing the production of a clean data file during data entry.

Knowing that the survey processing would revolve around intelligent data entry methods permitted DHS to develop a self-coding questionnaire. By carefully laying out the responses to questions and the boxes to code open-ended questions, DHS processing has not had to perform time-consuming and error-prone data transcription. Data entry operators can easily follow the flow of the questionnaire.

Even the earliest surveys in 1986 depended upon careful selection of data entry operators since data entry included data editing when problems occurred. In almost every case, data entry operators have attended the interviewers' training because they had to have a good understanding of the questionnaire being fielded.

Because DHS uses the data entry activities as computer-aided editing, very little time is spent on preliminary office editing. The batches of questionnaires returned from the field are checked to make certain that none are missing. Within each questionnaire, only the order of entries in the birth, contraceptive and health histories are checked. With the exception of complex internal consistencies, all other editing is done during data entry.

The data entry programs for the earliest surveys were not able to control the structure, ranges, skips and consistencies as stringently as the surveys now in the field. In fact, even the earliest surveys processed entirely with ISSA did not make full use of ISSA's capabilities; neither structure nor internal consistencies were completely controlled.

Currently, all structure, range and skip controls are programmed for data entry. The majority of consistency edits are included with correction required, if the survey team evaluates their data entry operators highly enough. Otherwise, the consistency edits are included only as warnings, not requiring correction, but providing an opportunity to correct miskeyed values.

Secondary data editing for inconsistencies not corrected during data entry has generally been performed within one to two weeks after the entry of the PSU. Much less time has been spent locating problem questionnaires among a few hundred rather than thousands. On the average, fewer than two secondary editing passes have been necessary on the PSU data sets. Because ISSA protects the data file from range, skip and structure errors during update activities, corrections seldom cause new errors. DHS has experienced a considerable savings in time with this method compared to the multi-pass editing required during traditional processing.

Data entry for DHS commences shortly after the beginning of fieldwork -- generally within two to three weeks. With such an early beginning, feedback on problems encountered in questionnaires can be given to the teams in the field so that the quality of their work may be upgraded.

Data have been entered in small data files: one for each PSU. In that way, no major disaster could occur because of hardware failure or operator error. Data have usually been entered onto floppy diskettes; however, data entry onto hard disks or Bernoulli cartridges has also been used.

A considerable amount of training has been required beyond that given during interviewers' training. The DHS Data Processing Manual provides guidelines for use during data entry and editing. These guidelines have been reviewed with survey personnel before their use; any changes made to the guidelines due to country-specific implementation have been documented. The agreed-upon guidelines have then been reviewed with processing personnel so that corrections could be made in a standardized manner.

Processing personnel have been asked to retain the ISSA listings that detail keying and questionnaire errors in data entry and secondary data editing. These listings provide documentation to keying and editing problems that surfaced during a specific survey's processing.

Basic to the success of DHS has been the selection and training of capable data entry operators and supervisors. What may appear as terribly slow data entry due to their need to make decisions and corrections has led to much faster processing. Preliminary studies of the data quality are positive.

5. Computer-aided Interviewers

Using centralized data entry, DHS has shown that data can be available for analysis within one to two weeks after the end of fieldwork. Nevertheless, there is strong interest in a move to computer-aided interviewers, not necessarily as a time-saving approach but rather as a method to improve the quality of the data collected.

Late in 1987, DHS conducted a field trial in Guatemala using ISSA on laptop computers. The text of the complete DHS questionnaire was transferred to data entry screens. Interviewers were trained in the use of ISSA for entering responses to their questions. No major problems were reported; in fact, the interviewers were extremely positive in their comments.

During Phase 2 of DHS, another study of computer-aided interviewing is planned. The objectives of the study have not yet been completely defined. However, the investigation will focus on whether the quality of data can be improved through resolving inconsistent responses during the interview.

6. Conclusion

Microcomputers have shown themselves to be excellent survey processing tools, especially when coupled with a strong integrated processing system like ISSA. The Colombia DHS completed fieldwork in December 1986 and the Regional Population Center (CCRP) presented preliminary results at a January 1987 seminar. The Department of Census & Statistics in Sri Lanka presented the DHS final results at a seminar in November 1987 for their survey fielded in the spring of 1987. The Preliminary Report for the DHS in Ghana was published in August 1988, only 2 months after the Ghana Statistical Service had completed the fieldwork. The Demographic and Health Surveys Program has experienced this type of successful survey processing in many developing countries.

Figure 1

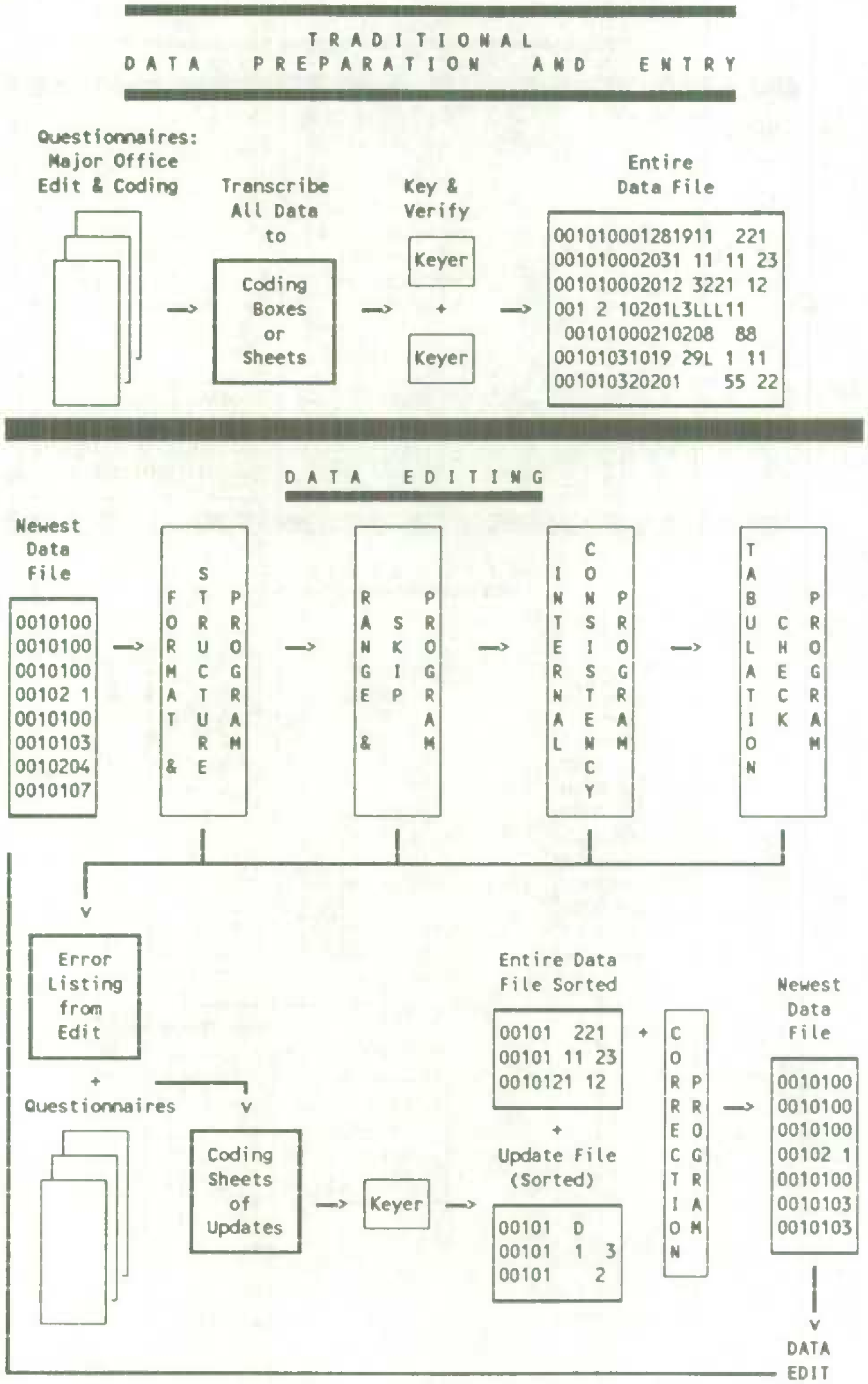
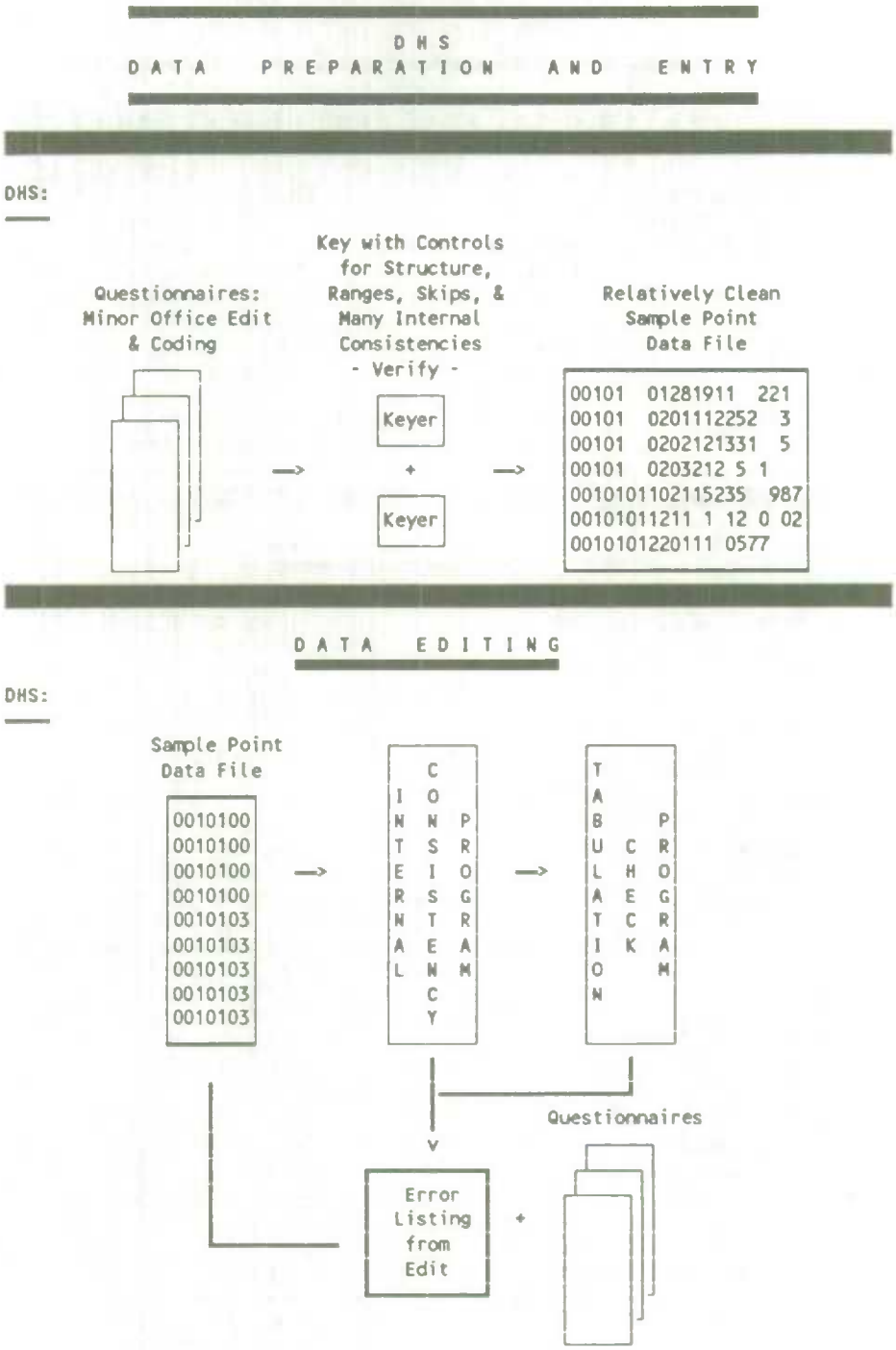

TRADITIONAL
DATA PREPARATION AND ENTRY
Questionnaires: Major Office Edit & Coding
Transcribe All Data to
Coding Boxes or Sheets
Key & Verify
Keyer
Keyer
Entire Data File
001010001281911 221
001010002031 11 11 23
001010002012 3221 12
001 2 10201L3LLL11
00101000210208 88
00101031019 29L 1 11
001010320201 55 22
DATA EDITING
Newest Data File
0010100
0010100
0010100
00102 1
0010100
0010103
0010204
0010107
FORMAT & STRUCTURE PROGRAM
RANGE & SKIP PROGRAM
INTERNAL CONSISTENCY PROGRAM
TABULATION CHECK PROGRAM
Error Listing from Edit
Questionnaires
Coding Sheets of Updates
Keyer
Entire Data File Sorted
00101 221
00101 11 23
0010121 12
Update File (Sorted)
00101 D
00101 1 3
00101 2
CORRECTION PROGRAM
Newest Data File
0010100
0010100
0010100
00102 1
0010100
0010103
0010103
DATA EDIT

Figure 2


DHS
DATA PREPARATION AND ENTRY
DHS:
Questionnaires: Minor Office Edit & Coding
Key with Controls for Structure, Ranges, Skips, & Many Internal Consistencies - Verify -
Keyer
+
Keyer
Relatively Clean Sample Point Data File
00101 01281911 221
00101 0201112252 3
00101 0202121331 5
00101 0203212 5 1
0010101102115235 987
00101011211 1 12 0 02
0010101220111 0577
DATA EDITING
DHS:
Sample Point Data File
0010100
0010100
0010100
0010100
0010103
0010103
0010103
0010103
0010103
INTERNAL CONSISTENCY PROGRAM
TABULATION CHECK PROGRAM
Questionnaires
Error Listing from Edit
+

A COMPARISON OF SOFTWARE FOR EDITING SURVEY AND CENSUS DATA

P. COTTON,

Statistical Consultant

A COMPARISON OF SOFTWARE FOR EDITING SURVEY AND CENSUS DATA

Paul Cotton

This paper describes a set of criteria which can be used to evaluate the capabilities of packages used to edit survey and census data. These criteria include: basic edit program functions, edit strategy, portability, level of integration, mode of operation and documentation. In addition, we examine advanced features such as integrated data dictionary, data entry program and tabulation capabilities. These criteria are then used to compare four new edit packages available for use on the IBM Personal Computer.

1. INTRODUCTION

In order to produce accurate results from a survey or census, the survey designer attempts to avoid, or at least minimize, errors in the data collected. Unfortunately not all sources of error can be eliminated and an important stage in the processing of every survey is the detection and correction of these errors.

Before the advent of first generation computers, most survey data error detection and correction was done manually either during the field operations or at the survey headquarters by vast clerical staffs. With the advent of powerful mainframe and mini-computers, many statistical agencies started to use custom computer programs to automate the detection of some of these errors while still depending on human intervention to decide on the type of corrective actions to be taken.

During the 1970's, custom computer systems began to appear that could automatically detect <u>and</u> correct some types of survey data errors. When confronted with the high development and maintenance costs of such systems, statistical organizations began to invest in the development of generalized computer edit packages which could be reused for different surveys (Graves 1976). Although several of these generalized packages or their direct descendents are still in use today, a new generation of editing packages has recently become available. These packages which all run on the ubiquitous IBM Personal Computer now present survey designers with the task of choosing which package, if any, is appropriate for their surveys' needs. This paper is an attempt to assist designers in evaluating which of these new editing packages might be appropriate for their census or survey.

Section 2 describes a set of characteristics that today's survey designer can use when evaluating an editing package. Section 3 then uses these criteria to compare the following four editing packages which are available for use on the IBM Personal Computer:

1. Generalized Edit and Imputation System (GEIS),
2. Integrated System for Survey Analysis (ISSA),
3. PCEDIT, and
4. Integrated Microcomputer Processing System (IMPS).

Section 4 presents a summary of the evaluation's results and conclusions.

2. CRITERIA FOR EVALUATING SURVEY AND CENSUS EDIT PACKAGES

This section outlines a set of characteristics which can be used to evaluate survey or census editing packages. These criteria will be used in section 3 to compare four recently released editing packages.

2.1 Basic Edit Program Functions

A survey or census edit package should include the following basic functions:

Input File Description
The edit package must have a facility for describing the layout of the survey questionnaire or data. This usually involves describing the layout of each "record" in the survey data file and the "fields" on each record. The description should also permit the specification of a field which determines the record type and of the "key" fields which uniquely identify each questionnaire. The edit package may also permit the user to specify how the key fields describe different survey strata (e.g. geographic areas).

Structure Checks
The edit package should be able to ascertain that all records of a particular questionnaire are present and that no extra records have been included. For example, in processing a population census the edit program should be able to determine which household and individual records belong to a particular questionnaire.

Valid Values Checks
The edit package should be able to determine if a field has a value that is inside or outside the valid limits for that field. These checks are concerned only with single fields or variables.

Consistency Checks
The edit package should be able to compare two or more fields in a questionnaire and to check their consistency. The fields may be in the same record of the questionnaire or in different records of the questionnaire.

Edit Rule Analysis
Some edit packages accept input in the form of a set of edit rules which express the validity and consistency checks for the survey questionnaire. Such systems usually include a program to determine whether the specified edits are redundant or contradictory and if any additional edits not specified by the user can be derived from the original edits.

Automatic Correction
The edit package should provide a facility for making corrections to data that fails either the validity or consistency checks. This can be done under direct control of the user's edit program, using such techniques as a "hot deck" or "cold deck", or automatically by the edit package. When the edit program carries out automatic correction it is usually based on a principle such as changing the minimum number of fields.

Corrected File
The edit package should create a file of corrected data. This can be a new file identical in format to the input file or the changes can be made directly to the original input file.

Extracted File
The edit package should permit the user to create a file that contains any combination of data from the original input file and any variables derived by the edit program. The user should be able to specify the format of this file e.g. ASCII data, SPSS/PC format, etc.

Reports
The edit package should be able to generate reports on the errors detected and how the data was corrected. These reports can include some or all of the following:

1. the number of times each edit rule failed,
2. the number of times particular values were assigned to a variable,
3. the edits rules that failed for each questionnaire.

The user should be able to select which edit reports are generated and the geographic levels at which the reports are produced. The user should be able to request a dump of all or some of the fields in a questionnaire.

2.2 Advanced Edit Program Functions

The basic edit program functions described above have been available for some time in several generalized edit programs which run only on mini-computers and mainframe computers. The new generation of integrated edit packages now available for the IBM Personal Computer combine these basic functions with a set of advanced functions which are described below:

Data Dictionary Program
The "glue" that holds an integrated editing package together is its data dictionary. The data dictionary contains the meta-data (i.e. data about data) for the designer's survey. A well integrated package requires that the designer describe this meta-data only once. This same meta-data is then used by each of the component programs in the integrated package. This is preferable to an environment in which the user must describe the survey data separately for each program.

The data dictionary contains information on the layout of the survey's data. Each of the survey's variables or any derived variables is given a logical name which can be referred to by each of the programs in the integrated package. The data dictionary should accept the definition of a missing value and a not applicable value for each field. The dictionary should also permit the user to define descriptive labels or names for the values of encoded variables and then to refer to these names instead of the actual codes in the integrated programs. These types of meta-data provide a user with a level of data independence which protects his investment from small changes in the formats of his data.

The data dictionary program should include a function to create an output file which includes all the survey's meta-data so that this can be used as part of the survey's documentation.

Data Entry Program

To permit a user to enter the survey data interactively on the IBM Personal Computer, an integrated edit package requires a data entry program. This program should be well integrated with the package's data dictionary and possibly with the basic error detection and correction modules of the edit program. The requirements of survey and census data entry programs can be summarized by the following list of major features (Diskin 1986):

1. programmable data entry screens,
2. ability to detect errors during input including validity checks and intra-field consistency checks,
3. questionnaire skip pattern definition and insertion of not applicable value for skipped fields,
4. automatic duplication and incrementing of fields,
5. programmable field entry sequencing (cursor control),
6. data verification capability,
7. questionnaire or record retrieval and modification,
8. speed of execution,
9. operator statistics.

Tabulation or Retrieval Program

If an integrated edit package includes a tabulation program, a survey designer will be able to use it before, during and after editing the survey data. If the package is missing this program then the survey designer will have to resort to using a separate program which will required the use of another data description language and another data manipulation language.

2.3 Integration

In today's micro-computer marketplace we are often forced to choose between products that exhibit different levels of integration. For example, we can either purchase a very good word processor, a very good spreadsheet program and a very good graphics program or perhaps one single package which integrates these three functions for a much lower total cost but which includes an average word processor, an average spreadsheet program and an average graphics program.

A well integrated package will present the user with a single consistent user interface across all functions e.g. LOTUS 1-2-3. For example, menus will have a similar basic appearance and function keys will carry out the same action in each program. Less integrated edit packages may use different user interfaces and therefore require more user training and will not be as easy to use.

2.4 Editing Strategy

A major decision in the design of every survey is the timing of the detection and correction of errors in the survey data. In some smaller surveys, it may still be possible to carry out manual edits of the questionnaires in the field or perhaps after the questionnaires have been returned to headquarters. As mentioned previously, this costly process is not an option for designers of most large surveys or censuses.

The survey designer must then decide how much editing of the data will be done during data entry and how much will be done after data entry. Unfortunately, this decision cannot be taken in isolation since the amount

of editing which takes place during data entry can have a great impact on the total time taken to enter the survey's data.

For example, the data entry program for a population census usually does only valid value checks on the individual data fields since any more extensive checking (e.g. consistency or intra-field edits) would require that the data entry operators correct more complex errors. This can have a large impact on the total time taken before the survey's data is ready for subsequent processing. On the other hand, in a small survey with a very large and complex questionnaire (i.e. many skips), the survey designer might choose to use an "intelligent" data entry program which will ensure during data entry that the each questionnaire had been completed correctly by the interviewer. In this latter case, the amount of data must be small enough that the survey designer can choose to slow down the data entry to ensure that each questionnaire has the correct structure. This is worthwhile since it is very difficult to have an edit program correct errors in such a questionnaire later in the processing.

The ideal edit package would permit the survey designer to select the amount of editing that occurs at data entry time and the amount to be done after data entry.

2.5 Interactive Versus Batch Mode of Operation

Most of the early generalized edit systems were batch oriented because they were developed in large mainframe computer environments which best supported this type of processing. Users of these systems usually had to code their input edit rules or edit program in a fixed format and then submit a batch job to the edit system for analysis. Even when the user had a syntactically correct program, batch testing of the program was still tedious and time consuming. It was only when the user had a "correct" program and began actually to edit his data that the batch oriented operating environment became less of a hindrance.

Today's low cost yet powerful personal computers have made interactive computing a reality. Personal computer users now expect a package to check the syntax of their input while they enter it and to be able to test their "program" at any stage of development.

With today's high cost of human labour and with the ever decreasing cost of computer power, the interactive capabilities of an integrated edit package are very important. On the other hand, designers of very large surveys or censuses would still like to be able to submit large batch jobs for execution. Therefore, an edit package which permits both interactive and batch execution will meet the requirements of more surveys.

2.6 Portability

Although today's micro-computers are very powerful, there are still some complex edit tasks which may be best suited for execution on a larger more powerful mini-computer or mainframe. A portable edit package that is supported on both micro-computers and mini-computers or mainframes would permit a user to develop a test application on a personal computer and then to transfer the edit program and/or data directly to a more powerful system for production runs. Due to the high costs of training users on different computer systems, an edit package with a portable user interface would greatly facilitate its use on different sized systems.

In today's rapidly changing computer environment, portability to different micro-computer operating systems such as OS/2, Microsoft Windows and UNIX may also be important to some applications.

2.7 Complex Data Structures

An editing package's ability to easily handle complex data structures can be the deciding factor in some selections. In particular, the designer of a household survey which collects information at both the household and individual level would want an editing package which can handle hierarchical data. Some survey questionnaires include a set of questions which are repeated a number of times e.g. a birth history in a health survey. A package's ability to easily edit a group of repeating questions would be very useful for such a survey.

For the survey designer with the task of entering and editing the data from a large questionnaire with a complicated skip pattern, a package whose data entry component facilitates the online editing of such a questionnaire would be of great use.

If the integrated package supports a tabulation facility then an ability to create special storage structures optimized for retrieval functions may be an important selection criteria.

2.8 Qualitative Versus Quantitative Data

Most household surveys and population censuses collect mainly coded or qualitative data while most economic surveys collect quantitative or numeric data. While general purpose edit programs can be applied to each type of data, some advanced forms of automatic edit detection and correction are oriented to handle either qualitative or quantitative data. The type of data to be edited will often be one of the most important criteria in selecting the best edit package. In fact, for certain approaches the survey designer may have to choose between a custom application and a specific generalized package.

2.9 Type of Input

An important characteristic of an edit package is the type of input that is required and therefore what type of user background is necessary to ensure the package is used easily and correctly. The type of the input is also important in determining whether users of the package will require specialized training.

Many edit packages are most easily used by programmers because their input languages are in fact specialized third generation programming language. Other packages accept more parametric input to ensure than users with limited programming experience can use the package. For example, several edit packages require only that a set of edit rules be specified. The rules are then analyzed and when appropriate a special program based on these rules automatically detects and corrects errors in the survey data.

2.10 Efficiency

Depending on the size of the survey and the computer resources available, the efficiency of a particular edit package can be a very important

selection criteria. For example, if an integrated edit package's data entry program cannot support data entry rates of over 10,000 keystrokes per hour then it would be difficult to use for a large survey or census in which the critical item on the project's overall schedule is usually data entry.

Special consideration should be given to the evaluation of the efficiency of edit packages that provide automatic correction of invalid data. Such packages, especially those that correct quantitative variables, are based on algorithms that can be very expensive to execute when the number of records, edit rules or variables is large.

Perhaps a more important efficiency concern is whether the user's edit program created by the edit package is interpreted or compiled. Although interpreted programs have less overhead when first created they generally run more slowly than programs that are compiled and then executed. A designer looking for an interactive edit approach will usually settle for a interpretive system while a designer looking for a batch oriented system for a larger survey or census will usually choose a compiled system. The documentation for the edit package should indicate its relative speed and if possible the machine readable example should include enough data to permit the package to be benchmarked in different environments.

The selection of an integrated edit package for use on a IBM Personal Computer will also depend on the type of computer it requires to be used successfully. The selection process should determine the amount of the following resources required by the edit package:

1. amount of random access memory (RAM),
2. amount of disk space required for the software, survey data and for temporary files during processing,
3. presence of a numeric co-processor,
4. special requirements such as "extended memory", "graphics monitor", etc.
5. minimum resources required for data entry program to permit inexpensive data entry stations to be created.

2.11 Documentation and Training

When selecting any software package the user must consider the form of the available documentation and any training courses that may be available. Although written reference documentation is important, for interactive programs running on an IBM Personal Computer the form of online help can be more important.

The documentation should contain the following items:

1. installation instructions and "getting started" information,
2. tutorial for novice users,
3. user's guide, and
4. reference manual.

The documentation for an edit package should include several different examples to acquaint novice users with the package's features. These examples should be available in machine readable form and if possible, the same examples should be included in the written documentation. The user's guide should also include a discussion on how the package can be applied to

different classes of problems.

2.12 Cost and Value

For some survey designers on a limited budget, an important selection criteria will be the cost of the package. This includes the cost of the edit package itself and any other specialized software packages which might be required to use the package. For example, an edit package may require a database package be purchased and learned to provide the underlying file structures used for the user's survey data. The evaluation should not only consider the cost in dollars but also look at the value of the software being purchased. A free software package that does not have adequate functionality has no value to a survey designer.

3. EVALUATION OF EDIT PACKAGES

3.1 Generalized Edit and Imputation System (GEIS)

GEIS (1988) is a numeric edit package currently being developed by Statistics Canada for use in processing economic survey data. GEIS will be used to impute data errors that remain after a preliminary edit and followup stage have eliminated all but minor and unresolved errors. GEIS is based on previous research carried out at Statistics Canada on the editing of numeric data (Sande 1979) and on the editing of coded data by Fellegi and Holt (1976).

This review is based on a copy of Version 4.2 of GEIS provided to the author by Statistics Canada solely for the purposes of this evaluation. The GEIS software is not complete and is not currently being distributed outside of Statistics Canada.

3.1.1 Overview

GEIS consists of a set of separate program modules that can be used by survey designers to edit their numeric data. Rather than develop one monolithic complex system, the GEIS designers have opted for a modular approach. This permits the developers to prototype individual modules, to enhance certain modules without affecting others and to add new modules in response to gaining experience with the overall system. Similarly, survey designers can choose to use only those modules which are needed for their particular survey needs.

GEIS is embedded in the ORACLE (1988) database management system (DBMS). GEIS assumes that the data to be edited already exists in an ORACLE database. Users of GEIS will therefore need to learn at least the Structured Query Language (SQL) used by ORACLE to create and manipulate "tables" of data in an ORACLE database. The ORACLE DBMS was selected to organize and handle the survey data and the GEIS control information because of its portability across a wide range of computer architectures. GEIS and ORACLE are currently being used at Statistics Canada on IBM Personal Computers, on a UNIX multi-user system and on a very large IBM mainframe.

The GEIS is currently planned to consist of the following six major modules:

1. questionnaire description,
2. quantitative edit description and analysis,
3. statistical edit description and analysis,
4. error localization,
5. imputation,
6. reports.

The survey designer first uses the questionnaire description module to inform GEIS of the ORACLE table or tables that are to contain the survey data. The survey data to be edited must be stored in a single ORACLE table but for monthly or annual surveys GEIS can also be told which tables hold the survey's historical data. The survey designer must then define at least one stratum for the survey data. Later processing of the survey data (i.e. error localization and imputation) is carried out on a per stratum basis. This permits a survey designer to develop different sets of edit rules for different subsets of the survey data.

All information provided by the survey designer and all results produced by the various GEIS modules are also stored in ORACLE tables. These tables are created during the installation of GEIS and can be examined at any time through SQL.

The survey designer then must specify the edit rules that will be applied to the survey data. Each edit rule is given a name for future reference and is expressed as a linear edit involving constants and "column" names used in the ORACLE table containing the survey data. For example, the following are valid GEIS edit rules:

```
0.5 * SALES  <=  EXPENSES
EXPENSES     <=  SALES
WHEAT + RYE + CORN  <  TOTAL - FALLOW
```

GEIS does not insist that the user express the edit rules in a canonical form but instead provides a program that automatically determines the canonical form of each edit rule. After the survey designer has entered the basic edit rules they can be grouped together into different edit groups which can be applied to different stratum of the survey data.

When the designer has a group of edits defined the next step is to have the edits analyzed. GEIS provides functions to check the edits for errors (e.g. column name not defined in ORACLE table), to generate any implied edits and to generate fictitious records which pass all of the edits but which lie at the extreme points of the acceptable region defined by the linear edits in the edit group. These last two items can be examined by the survey designer to ensure that the original set of edits were appropriate.

GEIS also provides a facility for outlier detection, referred to as the statistical edit module. This module permits a user to determine the upper and lower bounds of a variable or to determine the ratio of a variable's value to previous values. This module can be used to determine limits for the linear edits or to identify outlier data values for imputation.

When an edit group has been finalized it can be applied to a stratum of the survey data through the error localization module. For each record in the stratum, this module determines which edit rules fail and then determines

the minimum number of fields which must be modified to correct these errors. The survey designer can influence the choice of fields to be selected for imputation by supplying weights to the fields involved in the edit rules. When error localization is complete, the fields that must be imputed for each record are recorded in one of the GEIS control tables. GEIS also maintains information on which records passed all edits so that these records can be used later as donors during imputation.

GEIS supports three different types of imputation. The first permits a user to specify an imputation estimator which will determine the corrected value of a field based on one of a set of built-in algorithms. These algorithms can refer to a historical data value for the same variable or an auxiliary variable. Deterministic imputation can be selected by the user to impute fields that have unique solutions for each record in the stratum.

The third type of imputation is donor imputation. This module replaces invalid and missing values in a record from a similar clean record that is found using a nearest neighbour search. The similarity of the donor records to the recipient record in error is judged based on a subset of the correct, non-missing values. This module was not implemented in Version 4.2 of GEIS and therefore was not tested during the evaluation.

The final module of GEIS is the reports module. This module permits a user to generate seven different reports on the edits for a specific questionnaire. For example, this module permits the user to print all edits, a group of edits or to determine if any edits are currently not in an edit group. The system designer can also generate several other reports by using the ORACLE SQL facilities to print the contents of any of the GEIS control tables e.g. the results of error localization can be determined by printing the appropriate rows of the "fields to impute" table.

3.1.2 Evaluation

Basic and Advanced Functions GEIS implements the basic edit program functions directly through some of its own program modules and indirectly though some of the underlying facilities provided by the ORACLE DBMS. GEIS assumes that the survey designer has created a single ORACLE table for each time period of data to be processed. Thus GEIS, relies on the data definition facilities of ORACLE to create the input file description. If the user's survey data is not already in an ORACLE table or is spread across more than one ORACLE table then SQL or an ORACLE utility function must be used to create the single table image required by GEIS. GEIS does not directly include a facility for describing a multi-record questionnaire and therefore GEIS is not designed to carry out a structure check on the user's survey data.

GEIS does include the other basic features of an edit program: validity checks, consistency checks, edit rule analysis, automatic correction and a corrected data file. These facilities are oriented to processing only quantitative data. The user can use the ORACLE SQL facilities to create an extract file.

The reports available from GEIS are broken down by stratum and edit group. These reports are as follows: a) the number of times an edit rule failed, b) the number of edit rules that failed by questionnaire and c) the total number of edit failures and total number of records failing at least one edit. Other reports can be created by using SQL to process the internal

GEIS control tables. The user can dump a particular questionnaire by using the ORACLE SQL SELECT command.

GEIS depends on ORACLE to supply the advanced edit program functions: a data dictionary, a data entry facility and a retrieval program. The ORACLE data dictionary is used to define not only the user's survey data but also the control data maintained by GEIS for the user. Since the ORACLE meta-data is itself stored in ORACLE tables, a user can manipulate his own data, the GEIS control data and the system's meta-data with SQL.

As explained above, the mandate of GEIS does not include the need to support a data entry program since it is assumed that GEIS would only be used to impute data that had failed to be corrected during a preliminary edit at collection or data capture time. Nevertheless, a survey designer that must input data to test GEIS can use the SQL INSERT command to fill a table with test data. A user that must load a table with a large number of records would use the appropriate ORACLE utility program.

While the SQL language can be used to retrieve and display sets of records from one or more tables, its "tabulation" facilities are quite limited. Operations such as average, maximum, minimum and sum are supported but no built-in facilities exist for creating cross-classified tables.

Integration

GEIS exhibits a very high level of integration because it is embedded in the ORACLE DBMS which has been very successfully marketed on the philosophy of using the SQL*PLUS language and SQL*FORMS as consistent user interfaces. The MS-DOS version of GEIS evaluated by the author can be executed directly from the command line or through a menu system based on the SQL*FORMS utility. The menu system provides the novice user with an easy introduction to GEIS while the expert user can always revert to using SQL if direct access to the GEIS control tables is required.

Editing Strategy and Mode of Operation

GEIS is based on the philosophy of separating the editing of the survey data into a preliminary phase carried out during data capture and a second phase based on GEIS. Although GEIS when executed on an IBM Personal Computer appears to be a very interactive system, the system can also be used in a batch mode by direct execution of the GEIS program modules. This option is a strength of GEIS and should be maintained in the future.

Portability

The author only evaluated Version 4.2 of GEIS on an IBM Personal Computer. Inside Statistics Canada, this version of GEIS is also available on a Unix multi-user system and on an IBM mainframe. The existence of GEIS on these three different computer architectures can be mainly attributed to the portability of the ORACLE DBMS.

The GEIS modules are written in the portable C programming language and use the ORACLE PRO*C precompiler to permit the use of SQL to process the user's survey data and GEIS control tables. Thus GEIS should be portable to any computer environment which is supported by ORACLE and which has a suitable C compiler.

The menu system based on SQL*FORMS described above is also available on each of the above three computer systems. This portability of the GEIS

user interface will be important to survey designers unsure of which sized machine will be appropriate for their application.

Complex Data Structures

GEIS does not support complex data structures but assumes that the survey data to be edited has been placed in a single ORACLE table. If the survey data is already in a set of ORACLE tables then the SQL facilities to "join" and "project" different tables can be used to create a single table for GEIS. If the survey data is external to GEIS then a custom program may be required to create a "flat file" that can be loaded into an ORACLE table.

Qualitative Versus Quantitative Data

GEIS is oriented to processing only numeric or quantitative data.

Type of Input

GEIS accepts a set of edit rules which expresses the relationships among the numeric fields in the survey data. No programming instructions are required but the survey designer is responsible for developing a strategy for the order in which the data will be edited. The survey designer does this by splitting the survey data into sets of records (strata) and by processing different sets of variables (edit groups). The user can also choose to apply different imputation techniques to correct different variables. Combining these different processing techniques to produce error free survey data will require extensive cooperation between the survey's statistical and subject matter staff.

Efficiency

The largest unknown factor about GEIS is its performance. Although embedding GEIS in the ORACLE DBMS has many advantages which have been pointed out above (e.g. powerful data dictionary, portability), any application using ORACLE pays a certain price in performance for the generality provided by the DBMS. An unanswered question is whether this price will be too high to permit GEIS to be used in production survey work.

Several of the steps involved in using GEIS can be very compute intensive. For example, the algorithm used to generate fictitious records which pass all of the edits but which lie at the extreme points of the acceptable region defined by the linear edits can involve solving a large number of sets of linear equations. The current algorithm used by GEIS is based on the work of Chernikova (1965). If the user includes a large number of variables and a large number of edit rules in an edit group, the generation of the extreme points will be costly .

Another compute intensive operation will be donor imputation which was not available in the version of GEIS evaluated by the author. The documentation indicates that donor imputation will be based on constructing a K-D tree (Bentley 1975, 1979) to permit a nearest neighbour search to be used to find records which match the recipient record to be imputed. This operation's execution time will depend on the number of fields used to construct the tree and the number of possible donor records.

Documentation, Training and Support

A survey about to use GEIS must have access to an experienced ORACLE user. For example, just the initial installation of ORACLE and GEIS took the author nearly 10 hours because of technical difficulties that any user could encounter. GEIS assumes that the survey's data is already in an ORACLE table and although only a limited number of SQL commands are

required to create such a table, a novice user will find this a daunting task. While the use of ORACLE may present a very steep initial learning curve for a survey's staff, the advantage of this approach will be felt in the long term. Once the staff has learned how to use ORACLE, SQL and GEIS they should be able to apply this same knowledge to other surveys.

The GEIS documentation is inappropriate for use by a new GEIS user. The documentation evaluated by the author consisted of 11 separate technical documents, four of which described Version 4.0 and were therefore out of date. The author also had access to several research papers which included overviews of GEIS (Kovar 1988) and the underlying theory of numeric editing (Giles 1988). The GEIS documents contained no examples and no machine readable examples were provided with the software. No tutorial was available to introduce novice users to the system. The current user's guide explains what the user interface is instead of why and how the user should use it.

When further experience is gained with using GEIS with production survey data, the GEIS documentation should be modified to include advice on how the package should be used to solve different types of problems. This material should include some performance results to indicate the size of computer required for different applications of GEIS.

Cost and Value

GEIS for the IBM Personal Computer requires that the user have a license for ORACLE version 5.1A which costs approximately $1500 in Canada. The user's Personal Computer must be an IBM PC/AT or compatible and must have 640K bytes of random access memory (RAM) and at least 896K bytes of extended memory. ORACLE requires about 7.5M bytes of disk storage. GEIS requires another 1M bytes of disk space and in addition the user must have sufficient disk space for an ORACLE database to contain the survey data and the GEIS control tables.

GEIS is not currently being distributed by Statistics Canada. The system is still under development and its true value cannot be determined until it has been used to process several production surveys.

3.2 Integrated System for Survey Analysis (ISSA)

ISSA (1988) was developed by the Institute for Resource Development for the Demographic and Health Survey Program. This package was developed to assist in the processing of 35 large-scale demographic and health surveys carried out by the Institute in developing countries. The software was created to permit the three phases of survey data processing: data entry, data editing and data tabulation to be carried out on IBM Personal Computers with one single package.

This review is based on ISSA Version 1.0 and its user manual of approximately 300 pages.

3.2.1 Overview

When the ISSA program is executed, it presents the user with its main menu which consists of a list of choices on the bottom line of the screen. Each menu choice includes the number of a function key which can be used to select the desired program or action. The main menu includes the following choices:

1. help (not implemented in Version 1),
2. end the ISSA program,
3. dictionary editor,
4. batch processor,
5. dictionary report generator,
6. file manager (not implemented in Version 1),
7. generate export file for SPSS/PC,
8. generate dictionary cross reference listing,
9. forms editor,
10. data entry.

In starting a new application, the survey designer will first use the dictionary editor or forms editor to define a new data dictionary for the survey data. The fact that two editors are available is due to historical reasons and the documentation recommends that only the forms editor be used.

When defining a new data dictionary, the user is initially presented with the Dictionary Description screen where the first action is to define the location of the variables that are used to identify different "levels" or parts of each questionnaire. The user can describe flat files or hierarchical files with different "sections" (i.e. records). For hierarchical files, the user must also define the location of a field which will contain each section's identifier value.

The program that accepts the data dictionary information is a mode oriented editor which permits the user to work in Definition, List or Modify mode. The user can move between these modes by using a particular function key and each mode assigns a different set of meanings to the function keys. A menu of current function key meanings is always presented on the bottom line of the screen.

After the completing the Dictionary Description screen, the user is presented with the Section Description screen which permits the user to define the following information for the current section:

1. section name and descriptive label,
2. section code or identifier,
3. part or level of the questionnaire to which this section belongs,
4. whether the section occurs once or several times,
5. the minimum and maximum number of times a section can occur, and
6. the minimum and maximum number of times that grouped variables in this section occur.

As indicated by the above choices, the survey designer has a great deal of flexibility in designing how sections are used to contain the survey data. For example, ISSA permits the survey designer to model a set of variables which are repeated on a questionnaire in at least two different ways. Such a set of variables can be defined as part of a multiple section or record. The section would then be repeated once for each occurrence of the set of variables. The variables could also be defined together as an ISSA "group" and placed in a section which only occurs once but which permits the grouped variables to be repeated up to a maximum number of times.

After defining the section information, the user is prompted for information on each variable in the section with the Variable Description

Screen. ISSA permits the user to define the following information for each variable:

1. name and descriptive label,
2. location and length of variable within section,
3. format (numeric or alphanumeric),
4. number of implied decimals for numeric variable,
5. whether variable belongs to a ISSA group of variables,
6. a special value for missing data,
7. a special value for not applicable data,
8. default value for output variables,
9. valid ranges of values,
10. labels for each valid variable value.

When the user has defined the above information for the current variable, a function key is used to update the dictionary and ISSA presents the user with a new blank Variable Description Screen. When all the variables in a section have been defined, a function keys permits the user to start a new section or if all sections have been completed to exit from the dictionary definition program. The system designer can now use the dictionary report generator to create a hard copy listing of the dictionary information on sections and variables.

To develop an ISSA application, the user must add more information to a data dictionary. For example, to create a data entry application the user first defines a data entry form for each section of data in the questionnaire. by using the forms editor. Once the dictionary and data entry forms have been created, the user must then create ISSA "procedures" that will govern the data entry operation. Procedures can be added to the data dictionary by using the ISSA Full Screen Editor.

The ISSA command language is a structured high level programming language. There is no MS-DOS compiler or interpreter for the language and the procedure statements can only be executed by ISSA's data entry or batch processor. ISSA procedures can be defined for each level, section or variable in a survey's questionnaire. Each procedure can use the commands PREPROC (pre-processing) or POSTPROC (post-processing) to determine if it is to be executed before or after encountering the respective questionnaire level, section or variable. This type of specification permits the programmer to carefully control processing of hierarchical data files.

The ISSA procedure language includes the following types of commands:

1. assignment statements including use of arithmetic expressions,
2. a BOX command to permit recoding of variables,
3. program control commands such as:
   a) IF-THEN-ELSE-ENDIF
   b) SKIP TO
   c) WHILE-DO-ENDDO
   d) ENDLEVEL, ENDSECT and EXIT.
4. built-in numeric functions such as AVERAGE, COUNT and MAX,
5. built-in functions to process external files.

The above list is only a sample of the types of ISSA commands. A complete appreciation of the richness of the ISSA procedure language can only be obtained by reviewing the ISSA documentation.

Once the questionnaire's basic data dictionary has been modified to include the data entry forms and procedures then the data entry processor can be used to enter survey data. The same processor can also be used to verify data that was previously entered.

The survey designer can use the survey's basic data dictionary to create other ISSA applications. For example, a secondary edit could be created by taking a copy of the data dictionary and by adding new, different procedures that would be processed by the ISSA batch executor. The ISSA procedure language also contains a set of commands to permit the user to tabulate the survey data. A tabulation application can be created by adding procedures using these table commands to another copy of the survey's data dictionary and then submitting this job to the ISSA batch executor.

3.2.2 Evaluation

Basic and Advanced Functions

ISSA implements most of the basic edit package functions listed in section 2.1. The ISSA data dictionary can be used to define even the most complex survey data file and this description permits a ISSA data entry application to ensure that the structure of each input questionnaire is correct. Validity checks on individual variables can be determined using the data dictionary range specifications while the ISSA programming language would be used to test intra-field consistency. The ISSA programming language does not contain any special commands to automatically change incorrect data, but the language supports working variables and arrays which can be used to program custom solutions such as hot decks for imputation.

The basic edit package functions not implemented are edit rule analysis and edit reports. The input to ISSA is a programming language and as such edit rule analysis is not possible. No standard edit reports are available from ISSA but the types of reports described in section 2.1 could be programmed directly using the ISSA command language.

ISSA contains all of the advanced edit program functions described in section 2.2. A user can describe the survey data once using the ISSA data dictionary and then can combine a copy of this data dictionary with different forms and/or procedures to create the required survey applications. This requirement to copy the survey's data dictionary is a major problem with ISSA since a small change to the basic data dictionary would have to be replicated across all copies of the dictionary.

For some survey designers the richness of the ISSA data definition language will present a large number of difficult choices. Decisions such as how many levels and sections to define for a questionnaire and whether a multiple section or repeating group should be used for a particular set of variables must be made when the data dictionary is first created. These decisions are very important because they have a great impact on the form of data entry and how the ISSA programming language can be used in both the data entry and batch executors.

The ISSA data entry program is very powerful and includes all of the required features except operator statistics. One of the unique features of the ISSA data entry processor is its ability to maintain control when an operator wants to back through a skip pattern to change the path already selected. When this is done, ISSA guarantees that the appropriate

variables are set to the user defined not applicable code. Most other data entry packages permit an operator to modify any single variable on the screen without concern for what impact this might have on other variables.

The ISSA command language can be used to create tabulation applications and to create extract files based on the original data file.

Integration

ISSA's main design goal was to integrate the three main operations of survey processing. This goal has been carried out quite successfully. ISSA presents a consistent user interface across each of its program modules with most input provided by filling in fields and most actions specified by choosing an appropriate function key from an on-screen menu. When free format input is required to define forms or ISSA procedures, ISSA provides a specialized full screen editor to carry out the operation. Users can also create their ISSA procedures using their favorite DOS editor and then include these directly under ISSA control.

Due to the many different capabilities of ISSA, novice users may find it quite difficult to remember what the various functions keys do in each program module. Version 1 of ISSA does not contain an online help facility which could be used to solve this problem.

Editing Strategy and Mode of Operation

ISSA permits a survey designer to choose when the survey data is to be corrected. By adding sophisticated ISSA procedures to the survey data entry application, all data editing can be carried out during the original entry of the survey data. If the survey questionnaire is very complex and the data entry operators are experienced in using the questionnaire, then this can be the best method of correcting any errors. On the other hand, the survey designer can also use ISSA to create a data entry program that will only check that validity of individual fields. Editing of the resultant survey data can be carried out by creating a separate ISSA batch application.

Portability

ISSA is written in the C programming language (Kernighan 1978) for the IBM Personal Computer and is currently not available for use on any other computer or operating system.

Complex Data Structures

The ISSA data dictionary can be used to describe even the most complex survey questionnaires. Survey designers can break a questionnaire into different levels and sections (records) and can define repeating groups of variables. The ISSA programming language also permits the programmer to refer to external files which might have been output from a separate ISSA application. This permits a survey designer to split a survey's data into separate files but still be able to process the data together.

Qualitative Versus Quantitative Data

ISSA was originally written to process large scale health and demographic surveys in which the majority of fields are qualitative. The ISSA data dictionary permits the definition of multiple valid ranges for survey variables and the ISSA command language includes several commands specially oriented to processing qualitative variables. No special features are provided for processing quantitative data.

Type of Input

The design of an ISSA data dictionary and its associated applications for even a simple survey will have to carried out by experienced data processing staff. Although the forms editor is itself relatively simple to use, the decisions about how to implement a complex survey's questionnaire required the skills usually possessed by a systems analyst. The ISSA command language is also quite complex and would require a programmer to be used successfully.

Efficiency

Although the ISSA command language is interpreted by either the data entry or batch executor, the author's example applications ran very quickly. When an operator searches an ISSA created data file, finding a particular questionnaire is very fast since each file automatically has an index created for it. Tabulations on a large data file can be quite slow but the system appeared to be largely input-output bound. Only a different underlying data structure such as transposed files (Cotton 1979) is likely to speed tabulation applications.

Documentation, Training and Support

The ISSA Documentation Manual for Version 1.0 is over 300 pages long. This single manual acts as both an introductory user's guide and reference manual for ISSA. The manual is an adequate reference manual that could be used by an expert user since it contains a definition of all of ISSA's capabilities. The manual has a a master table of contents and each of the fifteen chapters is preceded by a detailed index. The evaluator would have preferred one overall and more complete index to assist in finding reference material on what particular functions keys meant or on the exact syntax of an ISSA command. More examples of using the ISSA command language are needed in the manual.

Unfortunately, the ISSA manual is not suitable for use by a novice user. Each chapter in the manual describes a particular part of ISSA in complete detail and this information is likely to overwhelm a new user. To further compound this problem many parts of the manual refer to a complex 13 page health survey questionnaire that is included in an appendix. A new ISSA user should be introduced to the package in stages and through much simpler examples that do not immediately require all of the richness of the ISSA package. A separate user's guide based on a simpler questionnaire and an online tutorial using this example would be extremely useful.

The ISSA system does not currently include an online help facility but the F1 function key is documented in all programs as providing this service. Novice users will find this facility very useful when it becomes available.

To further aid new users, ISSA should include machine readable copies of the examples used in the documentation. These examples should be complete and include not only an example data entry application and its basic data dictionary but also sample data files and the procedure and form descriptions for several ISSA applications.

Cost and Value

The ISSA software and documentation are available for about $500 in Canada. No other software is required to use ISSA on an IBM Personal Computer. ISSA provides very good value for its cost since it can be used for all stages of survey data processing.

3.3 PCEDIT

PCEDIT (1988) is a "intelligent" data entry and editing package for the IBM Personal Computer available from the Department of Technical Co-operation for Development, United Nations. This package has been developed under the auspicies of the Software and Support for Population Data Processing Project which previously distributed the UNEDIT, XTALLY and PC-COXTALLY statistical packages. These packages have all been developed in order to provide inexpensive solutions to processing population data, especially for users and statistical agencies in less developed countries.

This review is based on PCEDIT Version 2.11 and the documentation for Version 2.10 since the documentation update for Version 2.11 was not available.

3.3.1 Overview

The PCEDIT documentation states that it permits "intelligent" data entry. The package permits a user to describe a file layout for a questionnaire and a set of edit rules which PCEDIT uses to control the data entry, verification and modification of the questionnaire data. PCEDIT also collects statistics during data entry and verification of each batch of questionnaires.

All of the PCEDIT facilities are carried out through an interactive dialogue with the user. When PCEDIT is first invoked a main menu appears which permits the user to chose from the following seven functions:

1. File layout maintenance
2. Data entry program
3. Verification program
4. Modification program
5. Zero filling program
6. Data file information
7. System parameters

The file layout function permits the user to describe a new questionnaire or to modify an already existing layout. A PCEDIT questionnaire can consist of up to 64 different record types. Although there is no limit to the number of fields on each record, in practice the number of fields is limited by the fact that a record cannot exceed 80 characters.

When describing a multi-record questionnaire the field that determines the record type must be the first field on each record. Once this field has been described, PCEDIT displays the field definition screen with the the current record type, the current starting position and total record length at the top of the screen. The bottom line of the screen contains a status line with a short help message which indicates what the user should do next. This approach of using a status line to indicate the next action or current status is used effectively and consistently in all of the PCEDIT functions.

PCEDIT prompts the user for a field name, field type (alphanumeric or numeric), field length and verification status (YES or NO) for the current field. PCEDIT assumes each field starts immediately after the previous one so a field's starting position is determined automatically. Although a user

must define a record's fields in sequential order, errors can be corrected with some effort since a user can delete and add new fields.

After entering the descriptive field information the user can then enter up to 20 edit rules for the current field. These rules are based on a simple line-oriented key word language. When the user is ready to enter a new edit rule, the status line indicates all of the valid keywords. When the first character of a keyword is typed, the rest of the keyword appears automatically and PCEDIT prompts the user for the next valid keyword, operator or field name. This type of interactive rule development is only possible because of the rigid structure of the input language.

A field's edit rules can specify a field edit check which determines if the data entered is within a valid range of values and/or a consistency check with other fields already entered in the record. When particular values are encountered the user can specify that the data entry program request a new valid value, skip to another field, go to the next record or that an error message should be printed. Currently the user cannot specify structural level edit rules.

3.3.2 Evaluation

Basic and Advanced Functions

PCEDIT is not a complete editing package since it does not contain all of the basic edit programs functions (see section 2.1). In fact the package is really only a data entry package that permits a survey designer to describe the survey data file and to program valid value checks and consistency checks. This package does not currently support structure checks, edit rule analysis and the only file output by PCEDIT is the result of the data entry operation. No facilities are provided for carrying out automatic correction e.g. there are no built-in facilities for processing hot decks in the programming language. The only edit reports generated indicate the number of keying errors made by the operator and the number of corrections that were required.

PCEDIT contains most of the major features of a survey data entry package described in section 2.2. The greatest current deficiency is the user's inability to create data entry screens. PCEDIT currently only supports a line oriented data entry where the data entry operator's keystrokes are displayed on the screen exactly how they will appear in each record. PCEDIT does support the ability to detect input errors, to program a questionnaire skip pattern, to duplicate fields automatically and to program field entry sequencing. PCEDIT has a data verification capability and permits a user to retrieve and modify a questionnaire.

PCEDIT's data dictionary facility permits a user to name fields, describe their type and specify whether they should be verified. No facility exists for naming coded values.

Integration

PCEDIT presents the user with a consistent user interface and its highly structured menu system and context sensitive prompting make this an easy program to learn. The authors should try to maintain this consistency when they integrate more functions into PCEDIT.

Editing Strategy and Mode of Operation

A designer who chooses PCEDIT is selecting a package which only permits

interactive editing of the survey data during data entry or verification. No batch edit facilities are provided.

Portability
One of the main design features of PCEDIT was to ensure that it was very fast. To meet this requirement the package was coded in the IBM PC Assembly language. Therefore the package is not portable except possibly to another IBM Personal Computer operating system e.g. OS/2.

Complex Data Structures
As described in section 3.3.1, PCEDIT can process a questionnaire which requires up to 64 different record types. No facilities are provided for the definition or use of repeating groups of variables.

Qualitative Versus Quantitative Data
No special features are provided for processing quantitative data. The edit programming language does make checking for multiple ranges of valid values of a qualitative variable very easy.

Type of Input
PCEDIT has been written so that a non-programmer can describe the questionnaire layout and specify the edit rules. This is the major reason for the simplicity of the package's functionality and the use of the context sensitive prompting in all parts of the package. Whether end users will be successful at using PCEDIT will only be determined after the package has been used to process actual survey data.

Efficiency
Nearly all of the features of PCEDIT seem to execute instantaneously except for searching a data file for a particular questionnaire. This function can take some time and makes modification of particular questionnaires very time consuming. PCEDIT could benefit from the ISSA approach of generating an index file to speed random access to individual questionnaires.

Most data entry operators would expect to be able to used the numeric key pad of the IBM PC keyboard. Unfortunately PCEDIT uses the left and right arrow keys to move between fields. This should probably be changed so that the TAB and SHIFT TAB keys are used to move between fields so that the numeric key pad can be used exclusively for data entry.

Documentation, Training and Support
The PCEDIT software and documentation is currently available in English and French. The package also includes a demonstration program which takes the user through an example session of the file layout, data entry and verification functions.

The PCEDIT written documentation is very weak and needs a tremendous amount of work. The English manual is only 23 pages long, contains many grammatical and spelling errors but contains no examples or sample screens. The explanation of the edit rule syntax is very difficult to understand since the rules are described by the first letter of the keyword instead of by the whole keyword. The manual and demonstration program should be coordinated so that one consistent simple example is used in both. The file layout and some actual data files should be provided for such a sample application. If possible a separate more complex example could also be included for use by an expert user or data processing professional.

The written documentation should also contain an overview of how and why PCEDIT should be used. This material should include the steps to be carried out and what strategic decisions have to be made at each step.

The PCEDIT demonstration program is quite good but it is too short and does not contain enough examples of edit rules. The first screen of the demonstration indicates how to cause the demonstration to step forward and backward but does not mention if it is possible to resume the demonstration at any point other than the beginning.

The online help facilities of PCEDIT are one of its strengths. The user can always determine what functions are available by referring to the status line or by using a function key to cause a window to appear on the screen which lists the current actions of the function keys or CONTROL key sequences.

No formal support is currently available for PCEDIT although the program's author was very helpful during the evaluation of the package.

Cost and Value

PCEDIT is currently being distributed for free by the United Nations. This makes it an inexpensive choice for use as the data entry package for a census or survey application where multiple IBM PC's are going to be used. PCEDIT could easily support the high data entry rates expected in a census application i.e greater than 10000 keystrokes per hour.

PCEDIT could also be considered for the data entry and online editing of a complex survey questionnaire. If anything more than simple edits were required then the survey designer should consider a more complete edit package.

3.4 Integrated Microcomputer Processing System (IMPS)

IMPS has been developed by the International Statistical Programs Center of the U.S. Census Bureau. This system is based on the experience gained by ISPC on using micro-computers for processing census data in developing countries. The complete IMPS system as described by Toro and Chamberlain (1988) will include the following major components: a data dictionary (DATADICT 1987), a data entry module (CENTRY), an editing package (CONCOR 1988), a tabulation package (CENTS 1988), a census management and reporting system (dCONTROL) and a census planning package (CENPLAN). This evaluation will review only the first three of these modules that are required to enter and edit a user's census data.

This review is based on the following software for the IBM Personal Computer: a) DATADICT Version 1.0, b) a demonstration package for CENTRY and c) CONCOR Version 3.0.

3.4.1 Overview

IMPS has been developed over the last three years through ISPC's constant efforts to update and improve the software packages that it supports. In 1985, ISPC ported the CONCOR and CENTS packages to the IBM Personal Computer environment. At that time, these two packages required separate data definition stages although it was quite common for the output of CONCOR to be used directly by CENTS. At about the same time, ISPC carried out an evaluation of commercial data entry packages for the IBM Personal

Computer that could be used for survey or census data entry. Of course, use of one these packages required that the survey designer carry out a third independent and different data definition.

The introduction of a common data dictionary as defined by the DATADICT program has permitted the integration of CONCOR with the newly developed CENTRY data entry program. DATADICT can be used either interactively or in a batch mode on an IBM Personal Computer. The interactive mode prompts the user for the required information and produces a file of DATADICT batch statements which is then processed to produce a data dictionary file that can be read directly by CENTRY and CONCOR. If desired a user can create the DATADICT batch statements directly using a text editor.

DATADICT permits the user to define the layout of the survey or census data to be processed. When the interactive program is invoked the user is presented with a full-screen form which permits the user to define the following survey level information:

1. dictionary name,
2. record length of longest record,
3. default not-reported and not-applicable values which will apply for all variables.

The bottom line of the screen indicates what function keys or special keys are active. For example, by pressing function key 1 the user can receive online help on what the current screen is for. Online help is available for any specific field by pressing function key 2 after positioning the cursor in the appropriate field. The menu also indicates how the user can advance to the next screen or return to the previous screen.

Subsequent screens request that the user describe the position and type of a record type field (numeric or alphanumeric) and provide a name and value for each different record type. Then the user must define the "common" variables which occur on each record. These variables will usually be geographic or questionnaire identification variables. Finally the user defines the variables on each different record type. The information provided includes:

1. a long variable name from which an alternate short name is automatically derived,
2. the variable's length and starting position,
3. whether the variable is a subitem of a larger variable,
4. the number of times a repeated variable occurs,
5. the variable type (numeric or alphanumeric), and
6. the valid values and labels for these values.

When the user has completed the interactive input stage then DATADICT automatically generates the data dictionary file that can be used with CONCOR or CENTRY.

The CENTRY system permits a survey designer to develop a data entry application that permits the entry, verification and modification of census data and the collection of operator statistics. The CENTRY program consists of one module used by the developer to define the data entry application and another module used by data entry operators to execute the application.

The CENTRY developer module uses as input the data definition of the survey data file produced by DATADICT. The developer module permits the user to "paint" a data entry form on the screen using a full-screen editor. This module allows the user to design different forms for the batch information, for questionnaire level information and for each different record type in the questionnaire. The user can easily control the sequence in which the fields are to be entered and can cause automatic duplication of fields. The application designer does not have to define the valid values for each variable since these were already defined in the data dictionary. CENTRY does not support consistency checks between fields because these types of checks are considered undesirable for a census operation where the speed of the data entry operation is crucial.

The CENTRY execution module takes the application description created by the developer module and permits a data entry operator to create, modify and verify batches of survey data. The CENTRY execution module operates very quickly and can be used on an IBM Personal Computer with just 256K bytes of memory and two floppy disk drives.

IMPS depends on the CONCOR system to carry out the detection and correction of errors in the survey or census data. The CONCOR system has been used in over 100 different countries for the processing of population and housing data from national censuses as well as for agriculture censuses, labour force studies and educational surveys. CONCOR is written in COBOL and is available not only on the IBM Personal Computer but on a large number of mainframe and mini-computers.

Version 3.0 of CONCOR, which runs only on the IBM Personal Computer, uses a data dictionary file created by DATADICT to describe the input file to be processed. This data dictionary is combined with the user's CONCOR statements and if no syntax errors are encountered then a COBOL source program is generated which must then be compiled with a suitable COBOL compiler. The user's edit program reads a file of survey data, modifies the data according to the user's CONCOR edit instructions and produces an output file of correct data and any requested edit statistics.

The CONCOR language is a high level structured programming language created especially for writing computer edit instructions. The language consists of setup statements and executable statements. The setup statements permit the programmer to control the overall execution of the user's edit program and permit to user to define the following:

1. the data dictionary to be used,
2. the geographic area for which edit reports are to be created,
3. the set of variables which uniquely identify a questionnaire,
4. the maximum number of each record type expected to appear in a questionnaire,
5. the standard edit reports are to be generated,
6. the types of special checks to be executed e.g. zero-divide,
7. working variables that will used by the executable CONCOR statements.

The CONCOR language has a large number of different executable statements but they can be roughly divided into the following categories:

1. condition testing,
2. data assignment or transformation,
3. data imputation,
4. program organization and control, and
5. output creation.

Each CONCOR program is broken down in three basic parts. The PROLOG routine is executed once at the beginning of the program, the EPILOG routine is executed once at the end of the program and a FILTER routine for each record type in a questionnaire. The FILTER routines are "automatic" loops and the executable statements in these routines are repeated once for each record of that type within the questionnaire.

3.4.2 Evaluation

Basic and Advanced Functions

IMPS contains the majority of the basic edit program functions described in section 2.1. The DATADICT module permits the user the define the layout of the survey data file and CONCOR implements most of the other basic functions including: validity checks, consistency checks, automatic correction, corrected output file, and edit reports. No edit rule analysis is supported since the CONCOR system accepts a programming language and not a set of edit rules to be checked. CONCOR cannot produce an extracted file since the output file must be the same format as the input file.

The DATADICT and CENTRY components of IMPS implement two of the advanced edit program functions. These two modules meet most of the required features except as noted below. Although DATADICT permits a user to define labels for the codes of qualitative variables, then labels cannot be referenced in the user's CONCOR program. CENTRY does not permit intra-field consistency checks since it is aimed at census application where speed of the data entry is crucial. CENTRY also does not support questionnaire skip patterns and the automatic insertion of not-applicable codes for skipped variables.

As evaluated here, IMPS does not have a tabulation or retrieval function. But of course, in a census application a survey designer might use the IMPS CENTS system to produce camera ready tabulations.

Integration

The components of IMPS must be said to be loosely integrated since they use different user interfaces and input languages. Each of the programs which runs as an interactive program provides online help with the same function keys but the current menu appears in a different location in different modules. This is not a serious problem but a more consistent interface would assist novice users in learning to use IMPS.

The integration of IMPS could also be greatly improved by providing one main menu system which would permit the user to move easily from one IMPS module to another.

Editing Strategy and Mode of Operation

The editing strategy supported by IMPS is based on the assumption that it will be used for very large surveys with large data entry requirements. CENTRY only permits range or validity checks to be made on single variables - all other editing must be made through batch execution of the survey's CONCOR edit program. While the DATADICT and CENTRY programs permit

interactive development, the CONCOR system is inherently a batch process. The user creates a CONCOR program with a text editor and then submits this text to the CONCOR system for analysis. When the user has a syntactically error free program, the CONCOR system generates a COBOL source program which must in turn be compiled and linked to create the user's executable program. Each of these steps can take several minutes for a large CONCOR program. The offsetting advantage of the compile-execute approach is that the resulting executable program is very fast.

Portability

The DATADICT and CENTRY packages are currently available only on the IBM Personal Computer. Although Version 3.0 of CONCOR is not available on larger computers, Version 2.4 which is very closely compatible with Version 3.0 is available on many mainframe and mini-computers. It is very possible to develop an IMPS application on an IBM Personal Computer and then to carry out the production editing using CONCOR 2.4 on a larger computer.

Complex Data Structures

DATADICT permits a survey designer to describe a flat file or a hierarchical file with up to 50 different record types. The only restriction is that each record have its record type in the same location and that each record be a fixed size. DATADICT also permits groups or individual variables to repeat so that they can be treated as arrays of variables by the IMPS modules.

Qualitative Versus Quantitative Data

Most of the executable statements supported by CONCOR are oriented towards supporting the editing of qualitative data. No special functions are provided for editing of quantitative data.

Type of Input

The DATADICT and CENTRY packages are quite simple to use. A subject matter expert with minimal data processing training but who is familiar with the survey or census questionnaire should be able to create an IMPS data dictionary and set up a CENTRY data entry application in less than a couple of days of work.

Developing the CONCOR edit instructions requires a person with a good programming background. But it should be pointed out that the overall edit plan and strategy is more important than the CONCOR statements. No amount of sophisticated CONCOR programming can replace a comprehensive and well thought out edit strategy.

Efficiency

CENTRY data entry applications run very fast and will have no trouble supporting very high data entry rates. Once the survey's CONCOR edit program has been written and tested, execution speed is very fast. The author benchmarked a sample census application at over 10,000 records per minute on a 20 Mhz IBM PC compatible.

Documentation, Training and Support

The IMPS documentation varies from one component to the next. The CONCOR documentation which consists of a user's guide, installation manual and reference manual is excellent. The user's guide is aimed directly at the needs of a novice user and includes a complete example and many different code fragments to highlight different approaches available to the programmer. A machine readable copy of the complete example is provided

with the CONCOR package. The CONCOR user's manual also includes several useful and well-written introductory chapters on edit requirements in general and a final chapter which includes the steps in developing a successful CONCOR program.

The DATADICT documentation consists of a 52 page manual which must serve as both a user's guide and reference manual. Although the interactive DATADICT program includes an online help facility, the written manual which does not currently include any screen layouts should be enhanced to include a sample session with appropriate screen printouts. The written CENTRY documentation was not ready for this evaluation.

Cost and Value

The IMPS modules are available from ISPS for a nominal cost. The CONCOR (and CENTS) IMPS modules require a license for the REALIA COBOL (1988) compiler for the IBM PC which costs approximately $1500 in Canada.

4. CONCLUSIONS

The survey edit packages evaluated in this paper indicate the impact of high technology on survey and census data processing. Each of these packages permits a survey designer to carry out some or all of the processing for a survey or census on an IBM Personal Computer. The ability to use the interactive and personal style of an IBM PC for survey processing is a major advance made only in the last two or three years.

Several factors have been critical to the development of this new generation of survey edit packages. The most important is the continual increase in computational power and simultaneous decrease in cost of personal computers. As indicated in the research of Toro and Chamberlain (1988), it is now possible to carry out a national population census in a country of less than 10 million persons using micro-computers. Just three years ago, most countries of this size were using mini-computers for both the data entry and processing of such a census.

Another important advancement has been in the field of software engineering. In recent years, software developers have been able to take advantage of new tools in developing their software packages. For example, three of the packages described in this paper are written partly or entirely in the C programming language. Although the C language has been available to UNIX programmers for nearly a decade, the recent arrival of compilers based on a standardized C language for nearly every operating environment has made this language the choice for today's professional software development. The efficiency and portability of C permits software packages to be developed for today's selection of operating environments and then to be moved easily to new environments as they become available.

The use of the ORACLE database product by GEIS is another example of the use of new standardized tools. While the SQL interface to relational databases has been available in research systems since the 1970's, commercial database products which support the same SQL interface across a complete range of computer architectures have only recently become available. These new database products permit software developers to use them as platforms for the creation of new portable application packages.

While today's survey editing packages have benefited from recent technological advances, there is still a lot of work to be done in the future. Many of today's packages used to edit qualitative data such as CONCOR and ISSA still require that a programmer write a procedural description of how the survey data is to be edited. On the other hand, the Statistics Canada CANEDIT system, which requires only a set of edit rules as input, can be used by non-programmers. It remains to be seen if technological advances will permit a portable and efficient edit package based on the CANEDIT approach to replace packages such as CONCOR and ISSA. It is also currently unknown whether the GEIS package evaluated in this paper will perform efficiently enough to justify its use in processing medium to large surveys.

REFERENCES

Bentley, J.L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. **Communications of ACM 18,** 509-517.

Bentley, J.L. and Friedman, J.H. (1979). Data Structures for Range Searching. **Computing Surveys 11,** 397-409.

CENTS (1988). Census Tabulation System. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington, D.C. 20233. PC Release 5.0.

Chernikova, N.V. (1965). Algorithm for Finding a General Formula for the Nonnegative Solution of a System of Linear Inequalities. **U.S.S.R. Computational Mathematics and Mathematical Physics 5,** 228-233.

CONCOR (1988). Consistency and Correction Edit and Imputation System. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington. PC Release 3.0.

Cotton, P., Turner, M.J. and Hammond, R. (1979). RAPID: A Database Management System for Statistical Applications. 42nd Biennial Congress of the International Statistical Institute, Manila, December 1979.

Diskin, B. (1986). An Evaluation of Microcomputer Data Entry Software Packages For Use in Developing Countries for Census and Survey Processing. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington.

DATADICT (1987). Data Dictionary System. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington. Release 1.0.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. **Journal of American Statistical Association 71,** 17-35.

GEIS (1988). Generalized Edit and Imputation System, Research and General Systems, Statistics Canada, Tunney's Pasture, Ottawa, Ontario. Release 4.2.

Giles, P. (1988). A Model for Generalized Edit and Imputation of Survey Data. Business Survey Methods Division, Statistics Canada.

Graves, R. B. (1976). CAN-EDIT - A Generalized Edit and Imputation System in a Data Base Environment. The Electronic Data Processing Working Party of the Conference of European Statisticians, Geneva, March 1976 (available from Statistics Canada, Ottawa).

ISSA (1988). Integrated System for Survey Analysis, Institute for Resource Development, Westinghouse, PO Box 866, Columbia, Maryland 21044. Release 1.0.

Kernighan, B.W. and Ritchie, D.M. (1978). **The C Programming Language.** Prentice-Hall, Inc. Englewood Cliffs, New Jersey 07632.

Kovar, J.G., MacMillan, J.H. and Whitridge, P. (1988). Overview and Strategy For the Generalized Edit and Imputation System. Business Survey Methods Division, Statistics Canada, Ottawa. Working Paper BSMD-88-007E.

ORACLE (1988). Oracle Corporation, Belmont, California. Version 5.1A.

PCEDIT (1988). Software and Support for Population Data Processing, Department of Technical Co-operation for Development, United Nations, New York, New York 10017. Version 2.1.

REALIA COBOL (1988). REALIA Inc., 10 Riverside Plaza, Chicago, Illinois 60606. Version 3.00.

Sande, G. (1979). Numeric Edit and Imputation. Bulletin of the International Statistical Institute. Proceedings of the 42nd Session, Manila, 455-463.

Toro, V. and Chamberlain, K. (1988). The Use of Microcomputers for Census Processing in Developing Countries: An Update. Symposium '88: The Impact of High Technology on Survey Taking, Ottawa, October, 1988.

SOFTWARE FOR SURVEY ANALYSIS

W. A. FULLER,

Iowa State University

Software for Survey Analysis

by

Wayne A. Fuller

Iowa State University

Abstract

Several analytic techniques for complex survey samples requiring extensive computations are discussed. Items for which programs have been developed include: (1) variance estimation for statistics such as regression coefficients, parameters of measurement error models, entries in two way tables and quantiles, and (2) estimation for small areas using components of variance models. Software is under development to modify data to protect confidentiality of reports of individuals appearing in microdata releases.

1. INTRODUCTION

This conference is devoted to the impact of technology on surveys. I will report on computer software that has been developed for the analysis of survey data and that is under development for disclosure avoidance.

Some speak of the computer "revolution". I feel that a process that has been occuring relatively smoothly for all of my statistical career is better termed an evolution. We can now hold more computing power in our hands than was available to an entire university 30 years ago. Hence we routinely perform computations that we would not have considered in 1958.

One such area of computations, and one of the areas we will discuss, is the compution of variances and other complicated statistics for sample survey data.

The computer also creates problems for statisticians. Users desire their data in computer accessible forms. Because users now have computing power necessary for large data files, they request access to the microdata. These users have computing power for analysis, but they also have the power and software to perform matching operations. This creates problems for the statistician attempting to protect the confidentiality of respondents. We are currently working on a program that adds error to the observations to reduce the risk of data disclosure.

2. DATA ANALYSIS

In this section we discuss some personal computer software developed for complex survey samples. The basic program is called PC CARP and was designed for the IBM PC. The program was released in December 1986. This program evolved from the main frame program called SUPER CARP, developed at Iowa State University in the 1970's.

PC CARP is designed to compute estimated totals, ratios, means, quantiles, two-way tables, regression coefficients and coefficients of logistic regressions and to estimate the variances of these statistics. This program has been described elsewhere (See Schnell et al. (1988)). Therefore, I will only mention a few of the special properties of PC CARP.

The program is designed to run on a standard IBM PC with 450k of memory and a Math Co-Processor. Of course, it will also run on the newer more powerful machines. There is no limit on the number of observations or on the number of strata that can be handled by the program. PC CARP is menu driven. It is designed with short answers and default answers so that it remains fast to use for the accomplished operator. The program produces diagnostic statements, and contains "Help" and "Go back" options to aid the operator. All variance calculations are performed using Taylor approximations. See Fuller (1975) and Binder (1983).

The types of analyses available are given in Figure 1. This figure contains the menu from which the user selects the type of calculations to be performed. The questions at the bottom of the display appear one at a time after the user has selected a type of analysis. The questions permit the user to choose from the options available for that analysis.

```
                              PC CARP
                        Analysis Specification
                        ======================================
 Select the number corresponding to the desired type of analysis.  Replace
 // with the chosen analysis number (e.g. 08) and press the ENTER key.

                          AVAILABLE ANALYSES

          1. Totals                      7. Subpopulation Totals
          2. Ratios                      8. Subpopulation Means
          3. Difference of Two           9. Subpopulation Ratios
             Ratios                     10. Subpopulation Proportions
          4. Stratum Totals             11. Two-way Table
          5. Stratum Means              12. Regression
          6. Stratum Proportions        13. Univariate
                                        14. STOP Program Execution

          Key the numeric ID of desired analysis.......01

 Please respond( Y or N) to any additional question(s).
Press "ENTER" after each response:

Would you like covariances of the estimates to be computed?.....N
Would you like the design effect(s) to be computed?.....Y
```

Figure 1. PC CARP menu for analysis specification

Among the statistics produced by the univariate option are estimates of quantiles and an estimator of the standard error of the quantiles. The first step in the computation of quantiles is the construction of an estimator of the cumulative distribution function and the estimated variance of the distribution function. The quantiles and related statistics for a subpopulation can be computed by specifying a category of a classification variable.

The program is capable of constructing two-way tables and of calculating the test of proportionality for such tables. The program produces the covariance matrix of the proportions as an option. The

construction of estimates for subpopulations equivalent to one-, two-, three- and four-way tables is a particularly useful option.

The regression portion of the program provides the user with the ability to calculate residuals and to perform tests on subsets of coefficients. The standard output contains the coefficients and standard errors based upon Taylor approximations.

A supplement has been developed for PC CARP to compute variances for certain statistics for post stratified samples. The variances are for gamma post stratification. See Fuller and Sullivan (1987).

PC CARP requires a complete data set for analysis. For those desiring it, a hot deck imputation program, called PRE CARP, is provided with PC CARP. The hot deck operation replaces a missing value with the value for the same item from the record immediately preceeding the missing record that is in the same category of a specified classification variable. A special option in the program can be used to collapse a one unit stratum with the adjacent stratum.

PC CARP was designed to provide capabilities not readily availabe in other programs. Thus, PC CARP has essentially no capability for creating new variables from the existing data set. Programs such as PC SAS or PRODAS can be used for such operations.

A second member of the CARP personal computer family is the program EV CARP. The algorithms in this package are designed for the regression analysis of data observed subject to response error. It is the only package available capable of estimation for such regression models with data collected by complex surveys. The first algorithm of this type was written by Michael Hidiroglou in 1973. See Hidiroglou (1974) and Hidiroglou et al. (1976).

The estimation procedures require information about the nature of the response error in addition to the observations. The response error information may take the form of estimates of the error covariances, estimates of the reliability ratios, or instrumental variables. These estimation techniques are discussed by Fuller (1987).

EV CARP is a menu driven program patterned after PC CARP and the first set of menus for entering the data are essentially identical to those for PC CARP. Figure 2 contains the Analysis Specification display for EV CARP. The different types of analyses are associated with the different types of information that can be used with such models. The questions at the bottom of the display appear after the type of analysis

```
                    EV CARP
              Analysis Specification

Select the number corresponding to the desired type of analysis.  Replace
the / with the chosen analysis number and press the ENTER key.

                    AVAILABLE ANALYSES

        0. Weighted Least Squares     3. Minimum Distance (EV3)
        1. Error in Equation (EV1)    4. Instrumental Variables (EV4)
        2. Reliability Ratios (EV2)   9. Stop Program Execution

        Key the numeric ID of desired analysis........3

  Would you like sample summary statistics output (Y or N)?..........Y

  Taylor or Normal(SRS) Variance estimator (T or N)?................T

  Enter the value of the Alpha coefficient adjustment ( 0 or 1 ) ....0

  Would you like Residuals and Predicted values output(Y or N)?......Y

  Would you like to test coefficient sets equal to zero?(Y or N)?....Y
```

Figure 2. EV CARP menu for analysis specification

is selected and represent options associated with the selected procedure.

The output for the procedures is very similar to that for ordinary regression. The estimated coefficients and the associated standard errors are output. The program permits the user to test subsets of coefficients. Residuals and estimated true values for the variables subject to error are available as options in the program.

3. SMALL AREA PREDICTION

A program to compute predictions for small areas has also been developed at Iowa State University. The prediction procedure is based upon the components of variance model, sometimes called the empirical Bayes model. The model is closely related to some measurement error models and it is that relationship that is exploited in the program. See Fuller (1986).

The input into the program is the vector of direct survey estimates for the areas, any control variables for the areas, and the covariance matrices of the direct survey estimates. The current version of the program is rather small, and the total number of variables, the sum of the number of direct estimates and the number of control variables, must be no greater than ten.

The program computes a two round approximation to the maximum likelihood estimator of the between component of variance. This covariance matrix is restricted to be positive semidefinite and is used to construct predictions for the small areas. An estimator of the approximate covariance matrix of the predictors is also computed. Two forms of the covariance matrix can be computed. One form is an estimator of the unconditional variance of the multivariate predictor. The second form is an estimator of the conditional variance of the predictor, where the conditioning is on the observed vector being predicted.

The program is a modification of EV CARP and is menu driven in the same manner as are other members of the CARP family.

4. DISCLOSURE AVOIDANCE

The release of microdata files is becoming more common and the theory and software for matching records is continuing to improve. Thus concern for the confidentiality of respondents' responses is increasing.

One method of providing a level of confidentiality to respondents whose data appears in a microdata release is to add error to the responses prior to release. A program to add such error is currently under development at Iowa State University.

It seems desirable for the released data to resemble the original sample as closely as possible. The objective of the current program is for the released data to have a mean vector and a covariance matrix that are nearly equal to the mean vector and covariance matrix of the original data. Also, all univariate sample distribution functions of the released data will be close to the corresponding sample distribution functions of the original data.

The program proceeds in steps. The first step transforms the original observations into "nearly normal" variables using the inverse of the sample distribution functions. Normal error is added to the transformed variables, the sum is standardized to unit variance, and the standardized sum transformed to the original scale using the original sample distribution function. This program is being written in the matrix language IML of SAS.

ACKNOWLEDGEMENTS

Development of PC CARP was supported by agreement 50-YABC-5-66028 with the International Statistical Programs Center of the U.S. Bureau of the Census. Additional support was provided through Cooperative Agreement 68-3A75-8-12 with the Soil Conservation Service U.S. Department of Agriculture. Research on disclosure avoidance is being conducted under Joint Statistical Agreement JSA 88-1 with the U.S. Bureau of the Census.

The PC CARP and EV CARP personal computer programs were developed by W.J. Kennedy, D. Schnell, G. Sullivan and H.J. Park. Douglas Bearrood and Nancy Hasabelnaby contributed to the small area prediction program. Gary Sullivan is working on the disclosure avoidance project.

REFERENCES

Binder, D.A. (1983), On the variances of asymtotically normal estimators from complex surveys. International Statistical Review, **51**, 279-292.

Fuller, W.A. (1975), Regression analysis for sample survey. Sankhya C. **37**, 117-132.

Fuller, W.A (1986), Small area estimation as measurement error problem. In Proceedings of the Conference on Survey Research Methods in Argriculture, (D. Faulkenberry, Ed.) American Statistical Association and NASS, U.S. Department of Agriculture, Washington, D.C.

Fuller, W.A. (1987), Measurement Error Models. Wiley, New York.

Fuller, W.A. and Sullivan, G. (1987), Gamma Post Stratification. Report to the U.S. Bureau of the Census. Dept. of Statistics, Iowa State University, Ames, Iowa.

Hidiroglou, M.A. (1974), Estimation of regression parameters for finite populations. Unpublished Ph.D. thesis, Iowa State University.

Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1976), SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa.

SAS Institute Inc. (1985), SAS Language Guide for Personal Computers Version 6 Edition, Cary, N.C.: SAS Institute Inc.

Schnell, D., Kennedy, W.J., Sullivan, G., Park, H.J., and Fuller, W.A. (1988), Personal computer software for complex surveys. Survey Methodology To appear.

SUDAAN: A COMPREHENSIVE SOFTWARE PACKAGE FOR SURVEY DATA ANALYSIS

L. M. LAVANGE, B. V. SHAH, B. G. BARNWELL AND J. F. KILLINGER,

Research Triangle Institute

ABSTRACT

The development of a survey data analysis software package that is flexible enough to meet the needs of survey statisticians, as well as research scientists not necessarily schooled in survey sampling, is underway at the Research Triangle Institute. The system was designed to provide significant enhancements to RTI's existing software package, SUDAAN. The new SUDAAN system is written in the C language and provides greater efficiency and portability. The software development strategy consists of developing and debugging stand alone statistical procedures on an IBM PC with expanded capacity. The procedures are then implemented in two additional computing environments, VAX/VMS and IBM/OS. Variance estimation options and data analysis techniques not previously available in the RTI software package are among the many modifications incorporated into the new SUDAAN system. The purpose of this paper is to describe the design and development of this system.

KEYWORDS

variance estimation, user interface, C language

INTRODUCTION

Scientists at the Research Triangle Institute have been involved in designing, developing, and maintaining software systems for the analysis of survey data for the past 17 years. Recently, RTI has been under contract to the Public Health Service (PHS) to develop a comprehensive software package that meets the needs of statistical analysts at the National Center for Health Statistics (NCHS) and PHS. In particular, this package must be flexible enough to handle most of the statistical designs used by these agencies. For this purpose, we have embarked on the task of developing a system that incorporates many of the features of RTI's existing survey data analysis system but also includes significant enhancements. The purpose of this paper is to describe in detail the design and development of this comprehensive software package. The immediate ojective of the SUDAAN system design was to develop a series of procedures capable of performing statistical data analysis for complex sample surveys. The ultimate objective was to have a system that could aid in the design and evaluation of new statistical techniques.

RTI's existing software system consists of a series of procedures that produce statistical analyses in the form of weighted cross-tabulations, generalized ratio estimators, and linear and logistic regressions. All of the procedures were constructed using a unified set of FORTRAN subroutines that execute in the SAS computing environment on an IBM mainframe. In the course of developing a new system, we went through the process of rethinking our current design. Four goals emerged:

- Portability
- Reliability/numerical accuracy
- Computational efficiency
- Ease of modification/enhancement.

The system design for the comprehensive survey data analysis software, SUDAAN, evolved with these goals in mind. The first goal is being met by writing the new software entirely in the C programming language and testing simultaneously on several systems. Compilers for the C language are available on a variety of

operating systems. The initial programs are currently executing under the TURBO C compiler on an IBM personal computer, the SVC compiler on a 68020, and under the UNIX operating system on a GOULD computer. Versions are also being tested on the VAX/VMS and IBM/VS.

The second goal is being met by building checks into the system to assure that definitions of objects are meaningful and computations are feasible. We have implemented numerically stable procedures for the accumulation of sums of squares and cross products and for matrix operations, including matrix inversion.

For achieving the third goal of improved efficiency, special care has been taken in the design to isolate tasks that are repeated numerous times, such as a numerical calculation on each variable of each record. These sections of code are optimized to the extent possible without using assembly language. Some of the optimizing algorithms include processing of sparse matrices and vectors, replacement of multiplication by table lookup with addition, elimination of repetitive (in the loop) "if" tests by constructing separate do loops, and many others.

An important product of the SUDAAN design effort has been the creation of a high level statistical programming language. The SUDAAN language has been developed to provide the statistician/programmer with easy access to the software package, thus accomplishing the fourth goal listed above. Statistical procedures are currently being written in the SUDAAN language. Once they are finalized, some parts may be converted to C programs to maximize computing efficiency.

The SUDAAN design allows for four types of user interface as depicted in Figure 1. The general user will be able to access the system through user friendly procedures. Ultimately the procedures will be available through a statistical system, such as SAS, but the current design consists of stand-alone procedures that interface with a standard ASCII file format. The more sophisticated user can write SUDAAN programs for computations not available in the procedure library. This access feature is particularly well suited for testing modifications to existing procedures, such as the calculation of an alternative test statistic. A C programmer can access the system directly through C callable functions.

The key features of the SUDAAN system are described in detail in the following chapters. A brief overview of the software package for survey data analysis currently marketed by RTI is given in the following section. A description of the procedures being offered through the new software package, with emphasis on how the two packages differ, is also provided. The SUDAAN system design is the next topic. The SUDAAN language is illustrated with programming examples. The advantages and disadvantages of the SUDAAN design are the topic of the final section.

RTI SURVEY DATA ANALYSIS PROCEDURES

BACKGROUND

The collaboration of statisticians and computer scientists at RTI, beginning in 1971 (e.g., Folsom, Bayless, and Shah, 1971) and continuing today, has resulted in a series of programs for survey data analysis that reflects state-of-the-art

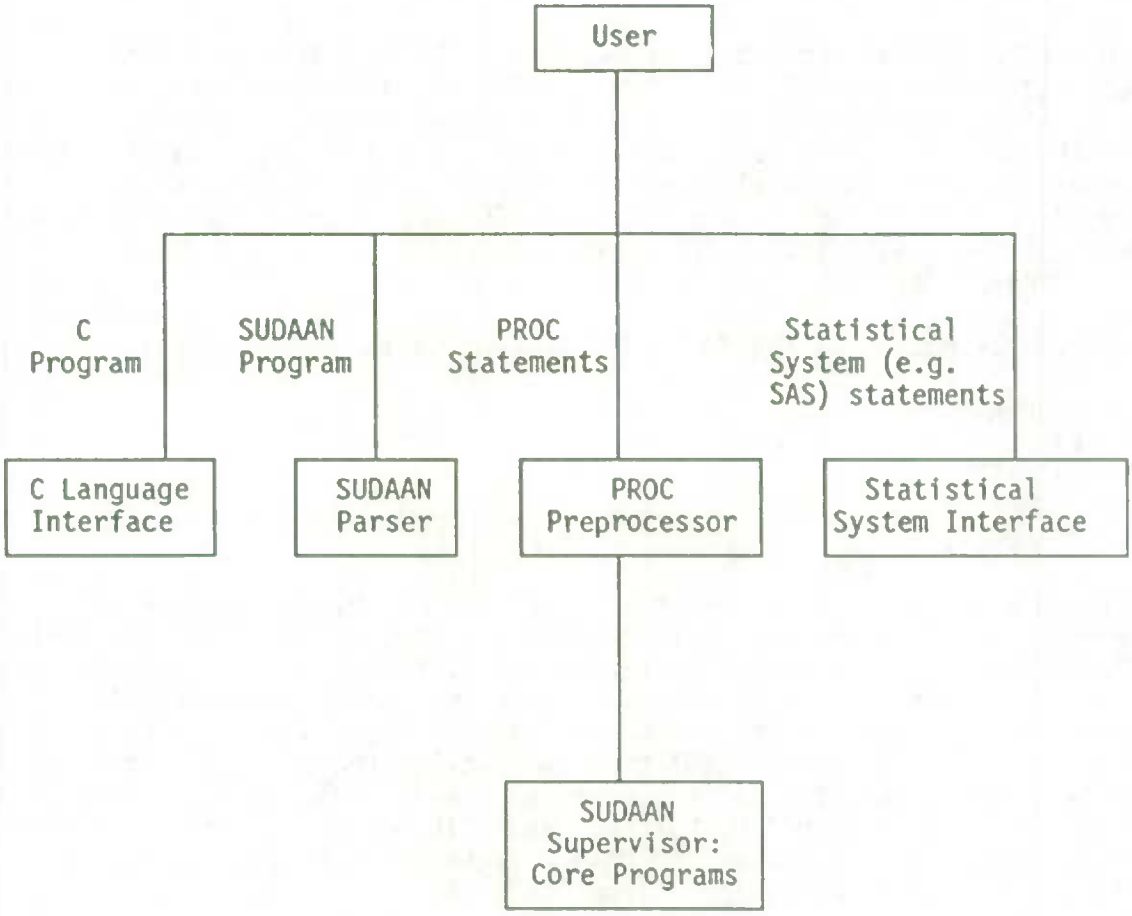


Figure 1. User Access to the SUDAAN System

methodology in this field. All of RTI's analysis packages employ the Taylor series or Delta method of variance estimation for complex sample survey designs (Folsom, 1974). Each program currently marketed is designed to run within the SAS framework.

Early work involved a variance component procedure developed to produce estimates of variance components associated with the strata and stages of an unequally weighted, stratified cluster sample. An updated version, VCMPNLS (Shah, 1971) became available as a SAS procedure in 1979. Weighted estimates of means, totals, and proportions and their standard errors were first provided by the program STDERR in 1974. A later version, SESUDAAN (Shah, 1981), enables the user to test for differences between domain estimates in a two-way cross-classification and to produce estimates standardized to a user specified population distribution. Modifications to SESUDAAN in 1980 and 1981 resulted in two procedures, RATIOEST (Shah, 1981) and RTIFREQS (Shah, 1982). RATIOEST produces estimates of ratios of two variables. RTIFREQS calculates weighted frequency distributions and estimates of row, column, and overall percentages for user specified cross-

tabulations. A further modification to RATIOEST that allows the user to input a sample unit level file and an event level file (e.g. medical visits with multiple records per subject) resulted in the RATIO2 procedure (Shah, 1982).

RTI's survey regression package, SURREGR (Holt, 1977), provides estimates of linear regression coefficients, their estimated variance-covariance matrix, and tests of hypotheses concerning the coefficients that are appropriate for a complex sample design. Logistic regression capabilities became available with the development of RTILOGIT (Shah, et al., 1984). This procedure runs in conjunction with SAS PROC LOGIST and produces pseudo maximum likelihood estimates of the parameters and tests of hypotheses about these parameters.

RTI is currently marketing the following procedures for general use:

- SESUDAAN
- RATIOEST
- RTIFREQS
- SURREGR
- RTILOGIT.

As mentioned earlier, variance estimation for all of these procedures is performed using the Taylor series linearization method. The RTIFREQS, SURREGR, and RTILOGIT procedures assume with replacement sampling, thereby enabling a first stage variance formula to be used. This approach is generally thought to be a conservative approximation for variances associated with more complicated designs. The SESUDAAN and RATIOEST procedures allow the user to specify with or without replacement sampling at each stage of a multistage design. Test statistics produced in SESUDAAN, SURREGR, and RTILOGIT are based on Hotelling's T^2 type statistic that is assumed to have a transformed F distribution in repeated samples (Shah, Holt, and Folsom, 1977).

New System Procedures

The SUDAAN system design consists of two phases of software development, each resulting in a series of statistical procedures. The Phase I procedures are primarily concerned with descriptive data analysis, including cross-tabulations, generalized ratio estimation, and quantile estimation. The Phase II design, currently in the planning stages, will incorporate several analytical procedures, including linear and log-linear modeling and survivorship analysis of survey data.

Several features are available under the new SUDAAN system that were not previously available with RTI software. Key among these are multiple variance options. Sampling variances are computed for one of three design options invoked by the user. These are:

- With replacement sampling at the first stage
- Simple random sampling at each stage (with or without replacement)
- Without replacement, unequal probability sampling at the first stage and simple random sampling (with or without replacement) at subsequent stages.

The second option requires the input of stratum specific population and sample counts for each stage. In addition to these counts, the third

option requires the input of joint probabilities of selection for each pair of primary sampling units within a stratum. These quantities are used to produce Yates-Grundy-Sen variance estimates for this design option. The three variance options cover a wide range of sample designs including sampling with probabilities proportional to size at the first stage. Unequal probabilities of selection at a subsequent stage can be handled by forming pseudo strata at that stage since probabilities of selection can vary from stratum to stratum. In addition to these variance options, the user may request that the variance-covariance matrix for estimates within a table be computed for all procedures allowing for cross-tabulations . Estimates of variance-covariance matrices were previously available only in the SURREGR and RTILOGIT procedures.

Phase I software consists of three procedures: CROSSTAB, DESCRIPT, and RATIO. The CROSSTAB procedure is the RTIFREQS equivalent in the new SUDAAN system. In addition to computing weighted frequency and percentage distributions, Chi-square tests of independence for a two-way contingency table are produced. The tests are computed using a Wald statistic according to methods advocated by Koch, Freeman, and Freeman (1976). During Phase II, the Rao-Scott Satterthwaite corrected test statistic (Thomas and Rao, 1984 and 1985) will be computed in addition to the Wald statistic for goodness of fit tests. This statistic has been shown by Thomas and Rao to have the best characteristics with respect to both power and significance level when compared with other statistics correcting for the liberalness of the Wald test.

The DESCRIPT procedure computes estimates of means, totals, proportions, geometric means, quantiles, and their sampling errors. Estimates can be requested for the subdomains of user-specified cross-tabulations. Options are also available to the user for standardized and/or post-stratified means and proportions and their sampling errors. The standardizing and/or post-stratifying variables and their distributions are required as input.

RTI has recently completed a simulation study comparing two methods for estimating the variances of quantiles for sample survey data (Wheeless, et al., 1988). The two methods considered were (1) the one proposed by Fuller and Francisco (1986) and currently used in the PC CARP software package and (2) a direct linearization method proposed by RTI. The simulation also compared two methods of estimating quantiles themselves. Based on the results of this study, the direct linearization method is currently being programmed for use in the DESCRIPT procedure.

The user can request estimates of linear contrasts among the domain estimates produced by DESCRIPT. Four types of contrast specifications are available. The user can specify a general linear contrast, all possible pairs of differences among levels of a domain variable, the differences between levels of a domain variable and the marginal value, and polymonial effects for an ordinal domain variable.

The RATIO procedure computes the estimates of generalized ratios for survey data. The user can specify numerator and denominator variables that are continuous or categorical. In the latter case, levels of the variables for which ratios are to be estimated are required. This procedure, as the DESCRIPT procedure, can be executed in one of three modes: basic estimation, standardized estimation, or post-stratified estimation.

The current plan for Phase II software development consists of four statistical procedures written in the new SUDAAN framework. The first of these is a general categorical data analysis procedure that will offer log-linear modeling capabilities for survey data analysis. This procedure will be a survey data analogue to the SAS procedure CATMOD. Procedures two and three will be linear and logistic regression procedures, developed based on the existing procedures SURREGR and LOGIST. Each of these will be reprogrammed in the new SUDAAN system and various enhancements added. Enhancements to the linear regression procedure are being planned for stochastic regression coefficients model analysis of survey data. Software has been developed at RTI under a contract to the National Institute for Child Health and Human Development for this problem (LaVange, et al., 1988) that will be incorporated into the SUDAAN system. Fourth, a general procedure for survival data analysis is planned that will allow the analyst to perform Kaplan-Meir estimation and Cox proportional hazard model estimation using complex sample survey data.

SUDAAN SYSTEM DESIGN

The immediate objective of the SUDAAN system design was to develop a series of procedures capable of performing statistical data analysis for complex sample surveys. The ultimate objective was to have a system that could aid in the design and evaluation of new statistical techniques. The system design for SUDAAN consists of three layers:

(1) PROCs or procedures similar to SAS or BMDP procedures
(2) SUDAAN data structures and functions operating on these data structures
(3) A system level interpreter for (1) and (2).

These layers correspond to different views of the system. SUDAAN may be perceived quite differently by different groups of users. In this section we discuss these views and describe in detail the major components of the SUDAAN system.

System Views

A substantive researcher or data analyst will view SUDAAN as a collection of statistical procedures for applying alternative statistical techniques for survey data analysis. A statistician may view it as a convenient tool for evaluating new statistical formulas or analysis techniques. SUDAAN allows for easy conversion of formulas into executable statements, thereby enabling new procedures or enhancements to existing procedures to be programmed quickly and efficiently.

A system designer may view it as a compiler, written in C, for a very high level language in order to specify statistical analysis tasks. For the computer scientist, the development effort consisted of designing an optimizing compiler or interpreter with the potential for global optimization of all available resources. The SUDAAN program defines only what is to be done, not how. The system is free to select any of the possible sequences of operations in order to achieve the user-defined result.

The PROC interface is in SUDAAN for the convenience of the user who repeatedly applies the same procedures to different datasets or different variables within the same dataset. The systems programmer's view represents one of the many possible compiler designs for the implementation of SUDAAN.

However, the primary force behind the SUDAAN design is the statistician's view, namely a tool for translating mathematical formulas into a computer program capable of producing numerical results for a given data configuration. The statistician need not specify how a particular estimator is to be calculated. Rather his focus is on what is to be calculated. For the statistician, SUDAAN is a "nonprocedural" language. Given the definition of a statistical quantity in terms of a mathematical formula, the system determines the computational algorithm. Users have no knowledge of the representation of the data structures in memory or on an input/output device and no control of the program flow. In this sense the SUDAAN language is viewed as definitional and not procedural.

The basic elements of the system are as follows:

- Statistical data objects and abstract data structures
- Functions of and operations to data structures
- Functions to input, output, and print data structures.

Each of these is described in turn in the following subsections.

Statistical Data Structures

An example of a statistical data structure is the dataset resulting from a national household survey. Such a dataset might consist of records containing information on a sample of individuals, each selected from a sample of households within enumeration districts, counties, and states. The actual dataset contains a record for each individual, consisting of various attributes and item responses for that individual and household. Each record is identified by several hierarchical identifiers corresponding to the stages of sample selection: state, county, enumeration district, household, and individual. An example of a different type of data structure is a summary dataset of county-specific population counts of individuals by characteristics, such as race and sex. Such a dataset is a collection of two dimensional tables or matrices with the two hierarchical identifiers state and county.

In SUDAAN, statistical data structures are defined as collections of multidimensional arrays with nested hierarchical identifiers. These data structures are referred to as Balanced Array Trees (BATs). Each of the datasets described above is an example of a BAT, the first with five nested identifiers and the second with only two. The SUDAAN language also consists of a series of functions and mathematical operators that are used to input and output BATs and to define BATs as functions of BATs and other input objects. The SUDAAN language itself is a mathematically closed set of BATs with respect to these functions. Applying a function to a BAT yields another BAT. The level of nested identifiers may change, but one or more BATs always result.

SUDAAN Program Examples

In order to illustrate the SUDAAN language concepts, consider the problem of computing the between primary sampling unit (PSU), within stratum mean square for an estimated population total. Let y(hijk) denote the observed values of the variable y collected from a nested sample design for the kth individual in the jth household, ith PSU, and hth stratum. The filename "EXAMPLE1" refers to an ASCII file with the nested sample identifiers included as variables one to four on the file. The variable y is the fifth variable on the file. The input

file is stored in a standard file format with a codebook describing the variables and associated levels. Variable labels or formats are automatically read in and used for printing. In the following statements, X refers to the BAT consisting of these data values. The program uses the summation function, SIGMA, to compute totals of y at various levels. The first argument of SIGMA refers to the BAT containing data values to be summed, while the second argument refers to the nesting level up to which the summing takes place. This argument equals the number of nested identifiers in the resulting BAT that contains the sums. Comments, delimited by the strings "/*" and "*/", further clarify the logic of the following program statements.

```
RECORDS = FICSFILE("EXAMPLE1",
          VECTOR(1,2,3,4));          /*RECORDS is a BAT containing the
                                       input dataset*/

X = SELECT(RECORDS,5);               /*X is a BAT of y values and id's */

PSULEVEL=2;
XSUM2=SIGMA(X,PSULEVEL);             /*Sum X over each PSU */
MH = SIGMA(CONSTANT(XSUM2,1),1);     /*Compute the number of PSUs
                                       per stratum by defining a vector
                                       of one's for each PSU in XSUM2*/

MHMIN1 = MH - 1;
STRATLEV = 1;
XSUM1 = SIGMA(XSUM2,STRATLEV);       /*Sum X over each stratum */
XSQR1 = SIGMA(XSUM2*XSUM2,
              STRATLEV);             /*Stratum sum of squares of
                                       PSU totals */

XSSO = SIGMA((MH * XSQR1 -
       XSUM1*XSUM1)/MHMIN1,0);       /*Compute sum of squared deviations
                                       over all strata */

MSE = SQRT(XSSO);
PRTABS(MSE);                         /*Print the result */
```

Since the SUDAAN language deals with symbolic definition only, the above statements can easily be modified to obtain the mean square errors associated with more than one variable. Replacing the second statement with the following results in the calculation of mean square errors for variables statement 5, 6, 9, and 10 in this example.

```
X = SELECT(RECORDS, VECTOR(5,6,9,10));
```

Next, consider the problem of computing mean square errors for estimated population counts corresponding to the cells of a multidimensional table. In the following example statements, TABLES refers to a BAT consisting of a two-way and a three-way table. The first is the cross-classification of variables 5 and 6 on the input file and the second is the cross-classification of variables 5, 9, and 10. The EFFECTS function converts categorical variables with specified numbers of levels into dummy (0,1) vectors. The CROSSP function defines the tables as cross products of these effect vectors. SIGMAWG returns two arguments, the unweighted and weighted sums for each cell in the tables. Once XSUM2 is defined, the program statements in the above example follow to compute the mean square errors. The required statements are:

```
RECORDS = FICSFILE("EXAMPLE1", VECTOR(1,2,3,4));
SUBGROUPS = VECTOR(5,6,9,10);      /*Variables defining subgroups */
LEVELS = VECTOR(7,2,2,2);          /*Associated variable levels */
OPTIONS = VECTOR(0,0,0,0);         /*Four options are off */
EFFVECTORS = EFFECTS(RECORDS,SUBGROUPS,LEVELS,OPTIONS);
TAB1 = VECTOR(1,2);                /*Variable 5 by variable 6 */
TAB2 = VECTOR(1,3,4);              /*Variable 5 by var 9 by var 10 */
TABVEC = VECTOR(TAB1,TAB2);
TABLES = CROSSP(EFFVECTORS, TABVEC);
WEIGHT = SELECT(RECORDS,11)        /*Variable 11 is the weight */
NWSUM = SIGMAWG(TABLES, PSULEVEL, WEIGHT);
XSUM2 = NWSUM(2);
```

Additional statements can be added to these examples to compute mean square errors for estimated percentages for each table. These require additional SUDAAN functions and are illustrated in the complete program example given in Appendix A. This program illustrates variance estimation for the CROSSTAB procedure assuming stratified simple random sampling with replacement.

SUDAAN Procedures

The task of developing new procedures in SUDAAN is fairly straightforward. The process involves writing a SUDAAN program and associating objects in the program with a well defined set of key words. The user-input to a PROC consists of a command level syntax, such as that used in IBM/TSO or SAS. The PROC processor interprets the user input and generates the necessary statements for the SUDAAN language parser. In short, writing a SUDAAN procedure is equivalent to defining a preprocessor for a SUDAAN program.

The major advantage of this approach is that the user has a SUDAAN program available for each procedure that serves as detailed documentation. The user can see the translation of input to output and the series of calculations that are performed once the procedure has been called. Modifications to the procedure can be made by modifying the accompanying SUDAAN program without writing a completely new procedure. The source for all of the SUDAAN procedures will be available to the user. An example is the source for the CROSSTAB procedure, also provided in Appendix A.

System Components

The system is composed of six major components:

- lexical analyzer
- syntax analyzer
- semantic analyzer
- executing supervisor
- function definitions
- PROC processor.

The overall system flow is depicted in Figure 2. As mentioned earlier, the entire system is written in the C programming language. The user-supplied

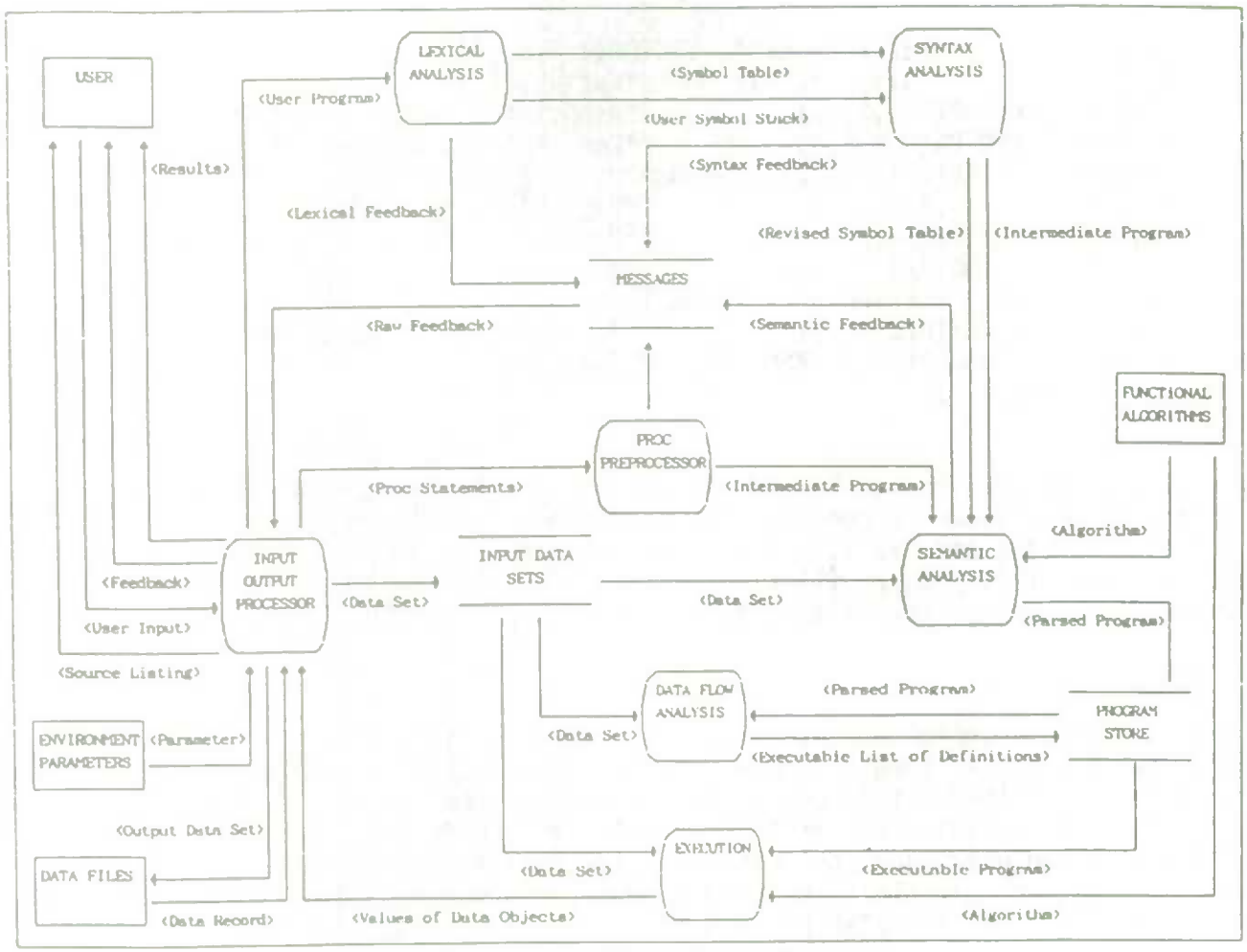


Figure 2. SUDAAN System Design

program is first passed to a lexical analyzer and a symbol table is output. This table is processed by the syntax analyzer into a list of executable statements. The semantic analyzer checks for valid function types, compatible arguments, and meaningful interpretation of the program statements. Error messages are supplied if necessary at this step. An intermediate output step is available to store the parsed program if so desired. An optimizing scheduler rearranges the order of execution of the program statements for optimal resource utilization. The executing supervisor then produces the program results. The input/output processer provides the appropriate user interface to the system. This may include a procedure interface that transforms procedural statements into a series of arguments for the SUDAAN program.

CONCLUDING REMARKS

The unique feature of SUDAAN is the entity defined as a BAT. Mathematically, a BAT represents a finite sequence of multidimensional arrays. The set of BATs represents a closed set with respect to binary operations and functions whose arguments are BATs. The functional syntax is "natural" for a mathematical statistician who can use the SUDAAN language to define the BATs of interest.

A SUDAAN user will have the SUDAAN source code available as part of the documentation. By referring to the SUDAAN program for a statistical procedure, the analyst will be able to determine exactly what underlying mathematical formulas are employed. When statistical research indicates new or improved formulas for a given analysis, the formulas can be readily translated into the SUDAAN language to yield a working computer program.

In writing a SUDAAN program, the user only defines the computational objects through function cells. The user does not need to specify how the calculations will be carried out. Increased reliability is achieved through checks built into the system to assure that the definitions are meaningful and that the objects required as output can be generated. The system is free to select the computational algorithm that is optimal given the dataset and available computing resources. The system also provides tremendous flexibility with respect to input and output. All input/output objects are defined as BATs. Any intermediate BAT can be saved or displayed at any time in the program.

The use of BATs, with the above stated advantages, has some restrictions and limitations. The user is limited to the BAT functions currently implemented, and all computations must be represented as BATs or functions thereof.

The SUDAAN programmer is confined to the available vocabulary of functions. Computations that require new functions of BATs may have to wait until such functions are implemented in SUDAAN. Over time, as more and more functions are implemented, this limitation will not be critical. The modular design of the system permits new functions to be added to the system without recompiling the entire system. Only the new function and its interface need to be compiled. It is also possible to test the new function independently before incorporating it into the system, thereby greatly facilitating the addition of functions to the system library on an as-needed basis.

Computations that cannot be represented as BATs will not be feasible in SUDAAN. The user must define all computations as functions of input BATs or data objects. The SUDAAN system is therefore not for general programming use, but should prove reasonably adequate for statistical analysis, and in particular, survey data analysis.

A major portion of the overall software development effort at RTI has been channeled into the SUDAAN system design. While this strategy has slowed the development of the statistical procedures initially, we are anticipating a tremendous savings in the long run due to the ease with which each procedure can be designed, implemented, and debugged. The end result of this effort will be not only a user-friendly software package that runs efficiently but also a valuable aid to the statistician for enhancing existing procedures and planning for additional procedures as new methodology becomes available.

REFERENCES

COX, D. R. (1972). "Regression Models and Life Tables" (with discussion), Journal of the Royal Statistical Society B, 34: 187-220.

FOLSOM, R. E., BAYLESS, D. L., and SHAH, B. V. (1971). "Jackknifing for Variance Components in Complex Sample Survey Designs." Proceedings of the American Statistical Association: 36-39.

FOLSOM, R. E. (1974). "National Assessment Approach to Sampling Error Estimation. Sampling Error Monograph. Prepared for the National Assessment of Educational Progress.

FRANCISCO, C. A. and FULLER, W. A. (1986) "Estimation of the Distribution Function with Data from a Complex Survey." Presented at the American Statistical Association Meetings, Chicago, Illinois.

HOLT, M. M. (1977). SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC 27709.

KAPLAN, E. L. and MEIER, P. (1958). "Nonparametric Estimation from Incomplete Observations." J. of the American Statistical Association, 53: 457-481.

KOCH, G. G., FREEMAN, D. H., JR. and FREEMAN, J. L. (1975). "Strategies in the Multivariate Analysis of Data from Complex Surveys." Inter. Stat. Rev. 43, 59-78.

LAVANGE, L. M., PFEFFERMANN, D., SHAH, B. V., and GABEL, T. J. (1988). "The Analysis of Relationships Between Growth and Dietary Intake Using the NHANES Dataset." Final Project Report. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B. V., HOLT, M. M. and FOLSOM, R. E. (1977), "Inference about Regression Models from Sample Survey Data." Proceedings of the 3rd Meeting, International Association of Survey Statisticians, New Delhi.

SHAH, B. V. (1981). RATIOEST: Standard Errors Program for Computing Ratio Estimates from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B. V. (1982). RTIFREQS: Program to Compute Weighted Frequencies, Percentages, and Their Standard Errors. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B. V. (1981). SESUDAAN: Standard Errors Program for Computing Standardized Rates from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B. V. (1978). "SUDANN: Survey Data Analysis Software." Presented at the Joint Statistical Meetings of the American Statistical Association, San Diego.

SHAH, B. V. (1979). VCMPNLS: Program to Compute Variance Components. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B. V., FOLSOM, R.E. HARRELL, F. E. and DILLARD, C. N. (1984): "RTILOGIT: Survey Data Analysis Software for Logistic Regression," Final Report, Work Assignment 74, Research Triangle Institute, Research Triangle Park, NC.

THOMAS, D. R. and RAO, J. N. K. (1984). "A Monte Carlo Study of Exact Levels of Goodness-of-Fit Statistics Under Cluster Sampling." Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC.

THOMAS, D. R. and RAO, J. N. K. (1985). "On the Power of Some Goodness-of-Fit Tests Under Cluster Sampling." Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC.

WHEELESS, S. and SHAH, B. V. (1982), "RATIO2 and RATIO3: Computing Ratio Estimates with Two Data Files. Research Triangle Institute, Research Triangle Park, NC.

WHEELESS, S. C., SHAH, B. V., and LAVANGE, L. M. (1987). "Results of a Simulation for Comparing Two Methods for Estimating Quantiles and their Variances." Project Report. Research Triangle Institute, Research Triangle Park, NC 27709.

APPENDIX A

SUDAAN Program Example

```
/* Example of a call to PROC CROSSTAB */

PROC CROSSTAB DATA=test120 DESIGN=wrs;
NEST stratum psu segno sampno;
SUBGROUP agecat race dsmsa dpovind;
LEVEL 7 2 2 2;
WEIGHT lwt;
TABLES race*dpovind agecat*dpovind agecat*dsmsa*dpovind;
```

C R O S S T A B

SUDAAN Program Statements

```
#include "constant.def"                /* file that defines constants */

filename = "test120";                  /* input data file */

nestvar = vector(1, 2, 3, 4);          /* position of stratum, psu, segment and person id on the input
                                          record. */

wgtvar = 5;                            /* weight variable position on input data record */

desnam = "WRS";                        /* design option equals with replacement sampling */

psulev = 2;                            /* position of PSU variable on NEST statement vector */

strlev = 1;                            /* position of stratum variable on NEST statement vector */

subgroups = vector(8, 9, 10, 11);      /* positions of AGE, RACE, SMSA, POVERTY STRATUM on input
                                          record */

levels    = vector(7, 2, 2, 2);        /* levels for each subgroup variable */

optionvec = vector (0, 0, 0, 0);       /* Option parms:  Include, Display, Nomarg and  Mergehi; */
                                       /* 1="Yes", 0="No". */
                                       /* Table definitions using Positions of variables in SUBGROUP
                                          vector *;

tab1 = vector(2, 4);                   /* RACE by POVERTY STRATUM */

tab2 = vector(1, 4);                   /* AGE by POVERTY STRATUM */

tab3 = vector(1, 3, 4);                /* AGE by SMSA by POVERTY STRATUM */

tables  = vector(tab1, tab2, tab3);    /* BAT combining three tables for processing */

records = ficsfile(filename,           /* input records with associated nest and weight variables */
          nestvar, wgtvar);

effvectors = effects(records,          /* create (0,1) indicator variables for subgroup levels or effects
             subgroups, levels,          vectors */
             optionvec);

tabvec = crossp(effvectors, tables);   /* form cross products of effects  vectors */

w = select(records, wgtvar);           /* input weights */

nsum0  = sigma (tabvec, 0);            /* compute sample counts */

wsum0  = sigma(w * tabvec, 0);         /* compute weighted counts */
```

```
21   rowper = margop(wsum0, wsum0, ROW,      /* compute row percentages */
            "/");

22   tcomplement = margop(tabvec,
                  tabvec, ROW, "-");

23   z = w * (tabvec * (1-rowper) +          /* Taylor series linearization individual-level */
            tcomplement * rowper)
            / wsum0;

24   zpsu  = sigma(z, psulev);               /* PSU sum of Taylor series deviations */

25   zstr  = sigma(zpsu, strlev);            /* Stratum sum of Taylor series deviations */

26   z2str = sigma(zpsu * zpsu, strlev);     /* Stratum sum of squared deviations */

27   npsu  = sigma(constant(zpsu, 1),        /* number of PSUs sampled within a stratum */
            strlev);

28   varwr = sigma ( (npsu * z2str - zstr    /* with replacement variance of row percentage */
            * zstr)/(npsu-1), 0);

29   serow = sqrt(varwr);                    /* standard error of row % */

30   print(vector(rowper, serow),            /* prints row % and standard error of row % for the 3 tables
          vector("Row %", "S.E. of row         defined in lines 11-13 */
          %"));
```

C R O S S T A B

Statistical Formulas for SUDAAN Program Example

1 h = stratum, $h=1,2,\ldots H$

i = PSU; $i=1,2,\ldots n_h$

j = sample unit; $j=1,2,\ldots n_{hi}$

w_{hij} = weight for individual (hij)

r = row of Table; $r=1,2,\ldots R$

c = column of Table, $c=1,2,\ldots C$

p_{rc} = row percent for the rc cell

23
$$z_{hij}(rc) = \begin{cases} 0 & \text{if sample unit hij is not in row r} \\ w_{hij} * (-p_{rc'}) & \text{if sample unit hij is in row r and column } c' \neq c \\ w_{hij} * (1 - p_{rc}) & \text{if sample unit hij is in cell rc} \end{cases}$$

24
$$z_{hi} = \sum_j z_{hij}$$

25
$$\sum_i z_{hi}$$

26
$$\sum_i z_{hi}^2$$

27
$$n_h$$

28
$$Var = \sum_{h=1}^{H} \left\{ n_h \sum_i z_{hi}^2 - \left(\sum_i z_{hi}\right)^2 \right\} / (n_h - 1)$$

29
$$Serow = \sqrt{Var}$$

CONTRIBUTIONS OF AI TO STATISTICAL DATA PROCESSING

J. H. JOHNSON,

Statistics Canada

Contributions of AI to Statistical Data Processing

J. H. Johnson

Statistics Canada
Research and General Systems
Room 2306, Main Building
Ottawa K1A 0T6

Abstract

Research into the subject of Artificial Intelligence (AI) has led to a number of advances in the exploitation of computer technology. Some of these may be useful for the acquisition, processing, and analysis of information performed by a statistical agency such as Statistics Canada. These contributions include the use of very high level programming languages, AI programming environments, the idea of knowledge representation, better man-machine interfaces, shallow reasoning, approximate reasoning, natural language processing, the AI workstation, expert systems and expert system shells. After an introduction to each of these, a brief discussion of its potential application to statistical data processing is provided.

KEY WORDS: *Artificial Intelligence; Applications; Statistical Data Processing; Expert Systems.*

Research into the subject of Artificial Intelligence (AI) [3,6,8,13] has led to a number of advances in the exploitation of computer technology. Ideas such as symbolic computation (the manipulation of symbolic as opposed to numeric information) and recursion as a programming technique have their origins in early work in AI and are now considered part of every programmer's set of tools. More recent work in AI has a similar prospect for providing techniques, approaches, tools, and ideas which will very likely be accepted ultimately in a similar way into mainstream computer science practice.

Unfortunately for the person unfamiliar with these new ideas, there is a cloud of confusion surrounding AI. The commentators on the potentials of the technology make claims which seem to be unsupportable based on the "conventional wisdom" of the computer professional. In addition, the enthusiasm of AI practitioners and researchers about their "toys" and what they can or will be able to do does not provide comfort to people more interested in "getting the job done".

What then are the expected benefits of this AI technology? To focus the discussion, we will consider the data processing and management needs of a statistical agency such as Statistics Canada. The term Statistical Data Processing (SDP) will stand for the Data Processing performed to support the acquisition, processing, and analysis of information as well as the access to information collected by such an agency. Of course, many of the comments will have a wider applicability, but this is accidental.

The remainder of this paper will give an overview of what, in the author's opinion, will constitute the main areas where advances in AI research will lead to benefits in SDP. These contributions include the use of

1. very high level programming languages (VHLL),
2. AI programming environments,
3. the idea of knowledge representation (KR),
4. better man-machine interfaces (MMI),

5. shallow reasoning,

6. approximate reasoning,

7. natural language processing,

8. the AI workstation,

9. expert systems (ES), and

10. expert system shells (ESS).

After an introduction to each of these, a brief discussion of its potential application is provided.

Throughout the remainder of this paper the phrase "general system" is often used. Where it appears it refers to a computer system which supports some aspect of the SDP process without being closely tied to a particular survey. It usually embodies a methodology which has evolved over time as a result of the successful application of techniques to particular surveys, as well as more complex methods requiring more investment than a single survey can justify. For example, the CANEDIT general system is capable of performing editing and imputation given only the layout of the input file and high-level "edit rules".

Very High Level Programming Languages

The programming languages usually used for AI work are often referred to as Very High Level Languages (VHLL). The thrust for research into VHLL arises not only from the needs of the AI community but from a more general drive to develop programming languages that are easy to use, and yet efficient, particularly on parallel architectures. There are three main approaches being investigated: functional programming, logic programming, and object-oriented programming.

Functional programming historically has included LISP[11], APL, and Backus' language FP. Recently a number of languages have sprung up as a result of an effort to provide a simple usable language free from side-effects based on the theory of the (typed) lambda calculus. One of the most well known of these is ML[14]. These languages have the advantage that they lead to extremely readable and writable code together with the prospect of compilation to a very efficiently executable form. They combine some of the best features of APL (high order abstractions close to the programmer's conceptualization) with the clear semantics of the lambda calculus and the readability of a language designed to be used.

Logic programming is an outgrowth of the success of Prolog [7,12]. Based on resolution theorem proving, it views computation as the process of proving the validity of a predicate by refuting its negative. The computation of results is achieved through bindings introduced by the unification process, that is, by the process of the implications of restrictions on the values that are made as the computation proceeds. New work is concentrating on extending the model while preserving the basic efficiency of standard Prolog.

Object-oriented programming views computation as associated with the objects on which they act. For example, to find out what value is associated with key 346 in a "dictionary" data structure, one sends the message `at: 346` to the data structure. The data structure uses its `at:` method to extract the value. Thus the nature of the algorithm used is localized to the data structure object. Objects are introduced into an inheritance hierarchy for methods. Thus the method for `at:` will most likely be really attached to a class that is a proper ancestor of the data structure class itself.

Another example involves the displaying or printing of objects. Every object can accept a message "print" in which case it displays its contents in some form to the user. Thus the knowledge about how to print things is decentralized (associated with the objects themselves) rather than centralized (grouped together in one giant print routine with a huge case statement).

A further discussion of VHLL with an example comparing the styles is provided in an Appendix.

VHLLs and SDP

Very high level languages provide a good potential for programmer productivity resulting from being able to use higher level abstractions, and declarative programming methods. There may also be more inclination for end-users to write programs themselves, bypassing system analysis and development steps and leading to a more experimental approach to survey processing.

In a way, however, this is simply continuing a trend that has been apparent for some time. General systems tend to naturally evolve their own command languages that have complexity matching their functionality. A system that controls edit and imputation must allow the specification of the edit rules and this is most naturally done using a declarative language designed for the purpose. In addition imputation must allow the specification of conditions under which a donor record is appropriate for hot-deck imputation as well as how the imputation is to be done. It is a relatively small step to allow escapes into a VHLL for the inevitable tailoring that must be done rather than a conventional programming language such as PL/1.

Artificial Intelligence Programming Environments

In AI work a very large variety of programming environments are in use. However, environments based on LISP and Prolog (or some derivative) are used for most work. Smalltalk, although it did not originate with AI has been used successfully for AI work.

These three programming languages share the view of computation as basically interactive and, as a result, environments usually include development tools to facilitate interactive development of interactive applications. Furthermore, all of the languages incorporate late binding, that is, delay the assignment of meaning to names until the latest time possible. Since early binding is what accounts for the efficiency of compiled code, many AI languages, by design, need to be interpreted instead of compiled. If they are compiled, the produced object code either blows up with many runtime checks or becomes simply a sequence of subroutine calls.

The need for efficiency has been recognized. This has resulted in compilers for LISP which, by bending the semantics a little and dissallowing certain programming techniques, can produce good code. There are native code compilers for Prolog which do very well. Prolog compilation is more feasible because it is a newer language with cleaner semantics. Smalltalk implementers have introduced some early binding on an *ad hoc* basis to allow better code to be produced by compilation. For example, the actual bindings associated with the True and False objects are ignored to allow for inline conditional tests.

These programming environments usually include a large library of routines to support interaction through a bit-mapped terminal at a high level, support high level file access, provide online editing of source or data, and other functionality as appropriate.

Although the availability of development tools and efficiency of delivered system affect the choice of the environment, the particular language used is also a factor.

LISP is an old language and has developed a large number of dialects. As a result, software portability between different LISP systems is difficult unless a very restricted subset is used. The problem is to some extent alleviated by the recent effort to standardize on a common dialect, appropriately named COMMON LISP [11]. Many vendors of LISP now claim to provide an implementation of this standardized language or, at least, are working toward compatibility.

Prolog is a much more recent contribution to the family of programming languages. Originally an attempt by Colmerauer and Roussel to base a language on the resolution theorem proving method of Robinson, it has rapidly developed into a full-featured programming language. Its use in AI results from its closeness to clausal logic and inference based on backward chaining since many AI researchers consider clausal logic as a good knowledge representation technique. It has advantages over LISP that unification and backtracking are an integral part of the language whereas they must be implemented by the LISP user often incurring a performance penalty. A Prolog with "tail recursion optimization" and a good garbage collector can emulate many of the useful features of LISP.

Smalltalk-80, although it has been around in some form for more than ten years, is only recently being advocated as an AI language. It is the prime example of the so-called object-oriented languages. AI people are interested in it because of the similarity to the AI knowledge representation technique of "frames". These, also called "slot and filler" systems, organize information into hierarchies where the child frame takes as defaults its parent values unless an explicit overriding value is given.

AI Programming Environments and SDP

Are AI languages and environments appropriate for SDP use? There are good prospects for them developing a niche although they will probably not displace "conventional" approaches for most software since they are currently quite expensive to run. In the longer term, as resources become cheaper, they will likely become more popular and will be introduced as expert systems based on them are introduced. They will also appear in applications that are not computer intensive such as the front-ends of general systems where the number-crunching data-intensive part might be implemented in a conventional language which gets access to its data through a database management system (DBMS).

Idea of Knowledge Representation

Another important idea which is basic to AI research is the notion of explicitly representing the knowledge inherent in an application. This contrasts sharply with conventional practice where the information is spread out in many different places and for the most part not even in machine readable form and often not even written down.

To make the point clearer, consider an application system of appreciable size (say of more than 20,000 lines of source). Ask yourself:

1. Where is the knowledge about how the system is organized and how it was designed, i.e, intended to be implemented?

2. Where is the knowledge about what was actually implemented?

3. Where is the knowledge necessary to run the system?

4. Where is the knowledge necessary to adjust the parameters of the system for optimum performance?

5. If the system is behaving strangely, is it possible to get an explanation of how it came to whatever conclusions that it did, and if so where is the knowledge which outlines this reasoning process?

For many systems the answers will be respectively:

1. The system organization and design is outlined in a specification document and/or a concepts document which may or may not be in machine readable form. If it is in machine readable form this is only for the convenience of text formatting; the machine cannot understand the document and comment in any intelligent way about its contents. Much of the important information may not have been written down at all because it was generally understood at the time or considered unimportant. The only source of that knowledge is the original design/implementation team although some of the information might be inferred by reverse engineering the source code.

2. The final arbiter of what exactly was implemented is the program itself in binary or source form. Not even comments in the source may correctly indicate what was implemented although they do provide information about what was intended or advice about how to look at a data structure or piece of code to understand its meaning.

3. The knowledge necessary to run the system is probably in a users' manual although this may be incomplete, out of date, or hard to understand. For any large system, pragmatically the knowledge about how to run the system is in the heads of the people that run the system. They know the special environmental conditions required for the system to run properly as well as features which do not work properly and require "workarounds".

4. The knowledge about the pragmatics of tuning parameters for optimum performance is distributed throughout the user community. If the system is widely used and complex enough, users may form users' groups to share their knowledge, perhaps even to the extent of organizing annual conferences, where novel ways of using the system are discussed. A number of "gurus" may appear who know the appropriate setting of parameters necessary for different purposes.

5. The knowledge necessary to have the system explain its behaviour is in the program itself (binary or source). Although this is machine

understandable in the sense that the sequence of operations are specified so that the machine can execute them, the only avenue open for the answering the question "Why?" is a low-level hardware trace as provided by general purpose debugging software or a higher-level software trace based on an instrumented copy of the source. Interpreting this information requires a detailed knowledge of the internal workings of the system for which the answer to question 2. must be consulted.

At this point it should be a little clearer what is meant by the word "knowledge", at least to an AI person. It is a symbolic representation which describes something sufficiently well so that the information can be used to define some process. The notion of process is important, in order to emphasize that more than memorization of some surface structure is going on. It is like the primary school teacher who says "Describe in your own words what the author is saying" or who provides exercises which require a student to look at a problem in a way different from the way it was presented.

For example, a database stores data, that is, numbers and text represented as patterns of bits. As the data is interpreted by a user or analyst it provides information. The number 25,309,330 by itself is data but when reported as the population of Canada [10] gives information. An example of knowledge would be a description of how this number is computed from other numbers, as is done in a spreadsheet.

Why is this view of knowledge important to AI? The answer lies in the observation that an important aspect of intelligence lies in the ability to understand and explain one's behaviour and reasoning. An artificial intelligence must therefore be able to explicitly represent and manipulate a significant part of its knowledge. Natural language understanding also faces the philosophical problem: what do you mean when you say that a machine understands something? The best answer is that the text has been translated into a form that contains exactly the information implicit in the text and yet can be manipulated by the machine in a rather direct way to draw inferences or to explain implied facts.

An interesting example of the AI approach to problems 1. through 5. discussed above is provided by the Smalltalk system.

1. The organization of the Smalltalk system [4,5] is provided as part of the structure of the database that contains the source and object (byte code) programs as it is presented to the user through the browser interface. This is basically organized as a 5-level [5] hierarchy: class category, class name, instance or class method, message category, and message selector. The browser uses a point, scroll, and click interface which allows the user to select an item at each level. After such a selection is made, the next pane provides a menu of possible selections for the next level. The last level selects the actual function to be displayed and a selection there results in the source appearing in a viewing window. Some of the levels are clearly for organizational purposes providing commenting information. At others, comments are used liberally.

2. When the user selects a source item by the means just outlined, he gets the current source: that is, the source of the currently running system. Thus the information about all aspects of the system are available directly to the user at all times. The source is all written in the Smalltalk-80 language, a very high level language.

3. The basic knowledge necessary to run the system is provided in printed form although it could easily be given as an online tutorial. Once a small number of basics have been learned, it then becomes easy to use the browser to obtain more advanced knowledge.

4. It is possible to build very general tools using the Smalltalk approach. These are shared through Smalltalk user groups.

5. The system can be made to explain its behaviour by means of various high level traces, or by directly modifying the source and observing the effect. Although dangerous (an unviable image can be produced), this approach provides a great deal of flexibility. It is also possible to view directly the state of any object by use of an inspector.

The main knowledge representation techniques are as follows:

- Logic: Since the time of Aristotle men have been trying to formalize language and reason-

ing in the interest of greater precision of thought. The result (Predicate Logic) has been an elaborate system capable of describing many aspects of the world. Knowledge can be represented declaratively by asserting as true a number of formulae. These are usually called axioms. Any formula which is true whenever the axioms are true, is then said to follow logically from the axioms. The system has a (sound and complete) proof system if there is a process to demonstrate by a sequence of small valid reasoning steps any formula that logically follows from the axioms, and none that does not.

- Frames: Another idea for organizing knowledge that dates back to the Greek philosophers is the idea of a hierarchical classification system. This is the idea used to organize the books in a library by discipline, and then by area within discipline, by sub area within area, and so on. In the AI context, a frame is a collection of "slots" each of which can be filled with a value or which reverts to a default stored in an ancestor within the hierarchy. Different systems allow a number of mechanisms for the filling of default values.

- Semantic Nets: A semantic net contains a collection of nodes representing concepts and arcs representing relationships between nodes. The main difference between this and the frame approach is that there is no requirement that the structure be a hierarchy.

- Rules: A popular way of representing knowledge in expert systems employs rules. These are the conditionals discussed under *shallow reasoning* discussed below and closely related to statements of clausal logic, and hence closely related to logic programming.

KR and SDP

The prospect of introducing knowledge representation techniques into SDP is quite bright. In addition to expert system applications, to be discussed later on, the idea of more completely recording knowledge about, for example, the conducting of a survey, in general or in particular, would assist survey control as well as facilitating the use of general systems. More specifically, a general system that understood the structure of the frame and survey design employed for a particular survey would be able to make better recommendations concerning estimation, editing and imputation, and confidentiality procedures than one that did not have this information.

Better Man-Machine Interfaces

Another of the features often associated with AI outside the research laboratory is the concern with the quality of the man-machine interface. This arises partly as a result of the complexity of the systems themselves and the resulting need for monitoring the behaviour of the system, as well as the targeting of the systems toward people not used to computers. Another aspect is the inherent nature of intelligence itself which involves also the communication of ideas, for example, by means of natural language.

A good example of the principles involved in good interface design is incorporated into the Apple Desktop Interface[1]. This document begins with the quote:

> "The Apple Desktop Interface is the result of a great deal of concern with the human part of human-computer interaction. It has been designed explicitly to enhance the effectiveness of people. This approach has frequently been labeled *user friendly*, though *user centered* is probably more appropriate."

It goes on to outline ten fundamental principles of the design.

1. "Metaphors from the real world": Use an appropriate metaphor with which the user will be familiar and will be comfortable. In Apple's case this is a desktop with folders that can be moved around, contain things, and can be opened to extract them.

2. "Direct manipulation": Actions from the user should cause observable effects to give the user the feeling of being in control.

3. "See-and-point (instead of remember-and-type)": The system should not require that the user remember anything that the computer already knows. It is much better to depend on the users ability to recognize. This is, for example, the principle behind menus but the

Apple Desktop Interface goes far beyond just using menus.

4. "Consistency": Things which do the same thing should have the same interface within and between applications. This implies, among other things, standards for interfaces.

5. "WYSIWYG (what you see is what you get)": There should be no secrets, magic, or hidden state information that the user must remember. The user should be able to see exactly what the effects of his actions are.

6. "User control": The user should feel in control. In particular, one should not get the feeling of being forced to wait to be asked something but can volunteer information as appropriate.

7. "Feedback and dialog": This principle requires that the computer tell the user what it is doing and communicate in the user's language.

8. "Forgiveness": Recognize the fact that users will make mistakes. Allow experimentation. If the user tries some action that is not reversible, indicate the nature of the danger and allow the operation to be backed out.

9. "Perceived stability": The interface should look the same over time. Things should be in the same relative locations.

10. "Aesthetic integrity": The display should look attractive.

These principles are elaborated in the Apple Interface Guidelines document[1] as they relate to this particular interface. In particular, details of the conventions which Apple applications should follow are provided so that implementors can provide consistency across applications. Adherence to this interface is considered important by users to the extent that reviews of software assess the degree of conformance to the standard as one dimension of quality.

Natural language interfaces are another aspect of better MMI that can be provided and are discussed in a later section.

Better MMI and SDP

Why do we want better interfaces to software? The simple answer is that people use programs and they are more effective at their job if programs are easy for them to use. In particular training costs can be reduced for a system that can teach concepts and commands as they are actually needed. Furthermore, the loss of time of people not remembering a needed command as well the opportunity cost incurred if a system is not used at all can be reduced.

Shallow Reasoning

An important principle in much AI work involves the notion of "shallow reasoning". Note that there is no derogatory connotation associated with this term. It merely describes the high level abstract reasoning which humans must do in order to successfully understand the world around them. Deep reasoning is, on the other hand, the kind of computation which computers can do well and with which humans have more difficulty: lots of computation using specialized algorithms.

Basically, reasoning is shallow if it is based on rules of thumb. Reasoning is deep if it is based on relatively certain "first principles" and computation of the implications of these in the given instance. As with many terms used in computer science, it is a relative term: reasoning can be shallower or deeper.

For example, consider a program which plays a game such as chess or bridge. A deep reasoning system would be primarily based on the rules of the game and would implicitly search the large tree of possible future lines of play: that is, the tree having the current position as the root, the positions accessible after the next move one level away, the positions accessible after the possible replies one more level away, and so on.

A shallow reasoning system would play the same way that humans do. A chess program would move pieces so as to systematically attack some part of the opponents defense, and would evaluate moves based on the effect they had on space, mobility factors, and material balance. A bridge program would bid primarily on point count.

Shallow knowledge (that is, the knowledge used in the process of shallow reasoning) is usually in the form of rules of the form:

```
IF Condition_1
AND Condition_2
AND Condition_3
AND ...
AND Condition_n
THEN Conclusion
```

The conditions identify the situation under which the "rule of thumb" is applicable, as well as providing "hedges" which further limit the rule to domains where its truth is more certain. Another form of hedge, discussed in the next section, addresses the problem that rules of thumb are hardly ever universally true.

Computation based on deep reasoning can always be put in the form of rules. However, this is not usually done since the fine structure of the problem usually suggests optimizations that are more direct if the problem is looked at another way. For example, the searching done by a chess program can be optimized by use of the alpha-beta algorithm, and massive improvements result from using the boolean and shift instructions of the underlying hardware for the generation of legal moves.

Interestingly, the best game playing programs are, for the most part, based on deep reasoning. The regularities and optimization opportunities available to deep reasoning systems seem to outweigh the cost of doing a brute-force search. This has been disappointing for many AI researchers who think that giving a program more knowledge will make it better than a dumb program that looks at everything, good or bad.

On the other hand, shallow reasoning triumphs when the "first principles" are not obvious, or too far from the problem at hand, or are incompletely known. Such is the case with medical diagnosis, that is, the identification of a disease or malady from its symptoms. Within relatively "narrow" domains of interest, systems based on shallow reasoning have been able to mimic the diagnostic capabilities of a human expert, leading to the notion of "expertise" as a body of mostly shallow knowledge empirically obtained, and to the idea of a "expert system" as a computer system which captures this expertise and the knowledge about how to apply it. This "meta-knowledge", or knowledge about how to use base knowledge, is important in that it aids in the efficient search of possibilities and provides an impression that the investigation is directed to a human using the system. Expert systems will be discussed in a later section.

The point here is that the choice of shallow or deep reasoning is an important one which will affect the performance of the resulting system dramatically. In cases where deep reasoning is feasible, it should be considered.

Shallow Knowledge and SDP

Shallow knowledge is not really new to SDP since edit rules, imputation techniques, record linkage rules and weights are all really a form of shallow knowledge. The experience of AI research with regard to the exploitation of shallow knowledge may be of great benefit to an environment where there is already much rule-based knowledge.

Approximate Reasoning

As mentioned above, rules of thumb are often "hedged" because they are not absolutely certain. They are only approximately true or usually true.

This was noticed in the development of medical diagnosis systems and led to the incorporation of approximate reasoning into the paradigm of expert systems. Basically the idea is to introduce a scale of certainty. For example, the value one might indicate certainty, zero absolute falsity, and intermediate values indicate intermediate levels of belief in the truth and falsity of the question. Or a "one" might mean certainty, a "zero" might mean a complete lack of knowledge, and intermediate values indicate intermediate levels of belief. In each system, each fact will have a value associated with it and each rule will, on application, cause values to be adjusted to account for the change in degrees of belief that the rule implies.

There are four basic approaches (with many variations):

- Certainty Factors: This is basically an *ad hoc* approach which was incorporated in the medical diagnosis expert system MYCIN to provide approximate reasoning. The value one indicates certainty, zero lack of knowledge, and minus one certainty about the falsity of

the fact. Rules also have certainty factors, and a system of formulae are used to combine the values yielding a result certainty factor. The approach is very simple to program and does not require much additional computation, but does not have a strong theoretical basis. It is not clear exactly what the numbers mean.

- Fuzzy Set Theory: The idea of fuzzy reasoning is based on the idea that in the real world facts and rules are true to some extent. This leads to the idea of indexing states of the world by numbers between zero and one indicating the degree of truth. The facts and rules with value one are completely (always) true. Facts and rules with value greater than or equal to 0.9 will be true in all models having value less than or equal to 0.9. In general, facts and rules with value greater than or equal to p will be true in all models having value less than or equal to p. Fuzzy models thus have a clear theoretical basis and semantics, do not require too many numbers, are easy to compute and stable under repeated rule firings. The main problem is that they might not correspond to the user's notion of what the numbers should mean.

- Bayesian Inference: The Bayesian approach view assigns a number in the range zero (falsity) to one (certainty) to each fact. This number is interpreted as a subjective probability and is manipulated using the calculus of probabilities. Numbers are assigned to rules as conditional probabilities, and evidence is combined using Bayes rule. This approach is firmly grounded in theory and quite understandable. Unfortunately, it does not capture the notion of ignorance. There is no way to indicate a lack of knowledge. It also is expensive and requires a lot of parameters to be set, many of which may be very difficult to properly assess.

- Dempster-Shafer Theory of Evidence: This approach [9] generalizes the Bayesian approach to address the problem of expressing a lack of knowledge. For the two hypothesis case, two values are expressed: p_t the subjective belief in the truth of the fact, and p_f the subjective belief in the falsity of the fact. It is not necessary that $p_t + p_f = 1$ although the sum must be less than or equal to one. The discrepancy $1 - p_t - p_f$ reflects the lack of knowledge with $p_t + p_f = 0$ indicating a complete lack of knowledge. The Dempster-Shafer approach requires more parameters than the Bayesian approach although some of the difficulties with the Bayesian approach are removed as a result of the possibility of expressing total or partial ignorance.

Approximate Reasoning and SDP

The use of approximate knowledge is implicit in the use of a sample survey and the recognition of sampling and non-sampling errors. The main benefit would be facilitating the incorporation of such partial knowledge into the processing and analysis based on this data. For example, the Fellegi-Sunter model [2] of record linkage is based on a hypothesis testing model close to the subjective Bayesian method.

A more radical suggestion might involve the use of edit rules which do not signal absolute failure but provide evidence of unusual or suspect conditions. As evidence accumulates that a field or collection of fields is implausible, an imputation could be requested. This could have the effect of providing more sensitivity to anomalies in survey data without forcing data into a straightjacket of tight constraints.

Natural Language Processing

Historically a very important component of AI research has involved natural language processing. This has included natural language generation and understanding as well as machine translation. All of these, if done well, would dramatically improve the accessibility of computer systems by allowing natural language interaction. In addition, a vast body of knowledge in machine readable natural language would become available in a new way: computers would be able to digest large bodies of information and filter and summarize it for humans with the inclination but not the time to be fully informed in our fast moving modern society. Furthermore, machine translation would break down language barriers without forcing everyone to communicate in English.

Unfortunately, this is all still a futuristic dream as much work needs to be done. It seems, for example, that to properly remove ambiguity

from most texts requires a lot of knowledge about the real world in general and the domain under discussion. Thus, a complete solution will probably require that some more fundamental and more difficult AI problems be solved.

What can computers do now? The answer is that they have been successfully applied in narrow domains of interest, particularly those where a specialized vocabulary with constrained syntax and precise semantics are in use. To some it seems strange that computers have less trouble understanding a legal document such as a will than, say, a novel.

There is also a serious danger in the use of natural language interfaces which any system using one must address. Since the computer's understanding of the human operator will be incomplete, it may jump to the wrong conclusions. There should be protections against the bad effects that this might imply. For example, dialogs should be arranged so that the information that the computer has inferred should be fed back to the operator for verification. Furthermore, this must be done in a way that will be natural and easy for the operator so he won't be tempted to bypass the verification. The information must have been extracted and processed. A simple syntactic transformation of the input will probably miss most semantic problems.

Another requirement is that the human operator should know the boundaries of the fragment of natural language that the computer understands. In other words, the operator must learn somehow how to communicate in the terms that the program knows. For example, a computer game may insist that the verb to use for acquiring objects be "take" so that one must type "take sceptre". The command "pick up sceptre" may not be understood. Note that simply adding "pick up" as a synonym for "take" does not solve the problem but puts it off. What about "acquire", "grasp", "grab", "get", "lift", etc.? A solution in this simple case may be as simple as a comment in a help page that lists the basic verbs allowed together with indications about how to find out more about the use of specific verbs such as "take".

It remains true, however, that programs that provide a natural language interface are friendlier and easier to use. People prefer to communicate with computers in a well-defined sublanguage of natural language than in a special purpose language designed by a systems analyst as a control language for his task.

NLP and SDP

There are a couple of ways in which natural language processing can assist SDP. The first, implied by the above discussion is as part of a better man-machine interface to general systems. Another direction for application is related to the acquisition of data itself. Much of the data obtained is most naturally expressed by the survey respondent himself using a few words, or a phrase. As long as this is simple, this may involve only a table lookup probably after a correction for spelling variations. However, as the texts become longer, the natural language processing technologies may become useful. As long as the program can understand a worthwhile fraction of the texts provided and knows when it doesn't understand (so that a human expert can be consulted), it may be cost-effective to capture the full text and analyse it later rather than clerically analyse the text on the questionnaire itself and capture only the clerically assigned code. This might also provide better uniformity over the whole survey as well as providing a better "audit trail" for information derived from the survey.

The AI Workstation

The particular needs of AI research have led naturally to the evolution of software and hardware to support them. In terms of hardware this meant high performance machines which can support the heavy requirements of, in particular, large LISP programs. These machines require a lot of primary and secondary memory as well as being quite fast.

Why do these LISP programs require so many resources to obtain good performance? The requirement for large quantities of main memory result from a basically profligate approach to data structures. Pointers are heavily used, and pointers require storage. A LISP list requires at least two pointers for each element in the list: one to point to the element and one to point on to the rest of the list. The use of data structures more appropriate to other languages can also lead to the need for extra pointers. The philosophy of allocating new storage rather than carefully reusing it and

leaving the task of reclamation to a "garbage collector" leads to a requirement for a lot of memory to avoid the repeated need for garbage collection.

The requirement for increased speed follows also from the above arguments. Pointers must be followed to find the data elements and this must be done every time the value is needed. Similarly, garbage collection can require a lot of time since it involves an "audit" of storage, determining whether each particular block can potentially be reached by the continuing program.

Another reason for the requirements of high power is that the AI software development approach makes use of prototyping with inefficient algorithms. This allows the programmer to concentrate on ideas rather than focusing on what can be efficiently implemented. There are both good and bad aspects to this but, at least, it encourages the postponement of the necessary real-world compromises required for efficiency until later in the development process, with the result that fewer of them are found to be necessary.

The necessity for the large amount of secondary storage follows from the need to store inherently large programs with large libraries together with the need to maintain knowledge bases.

An AI workstation usually provides a bit-mapped display terminal to facilitate high-level graphics interfaces. Since applications that involve the use of AI tend to require complex interactions with the system user, it is good to have as many interface techniques available as possible in order to make the system easy and natural to use.

The AI Workstation and SDP

The main impacts of the use of this technology are as follows:

1. Programmer productivity is improved through the better turnaround and better tools of a high performance environment.

2. User productivity is improved though use of more natural interfaces supported by increased power.

3. More data analysis is possible and easy to do with relatively cheap computing readily available. This could lead to a better understanding of our data and the improvements in published statistics that should follow from this.

4. An experimental approach becomes feasible when computing is cheaper allowing reprocessing using alternative methodologies, or replication to assess the magnitude of errors introduced by the methodologies used by general systems.

Expert Systems

Much of the early AI research was concerned with general approaches and general strategies. These were demonstrated to be ineffective in solving real problems since they were simply too expensive. Success was achieved when researchers focused on specific domains. The success of MYCIN in particular at diagnosing bacterial infections established the paradigm:

> Build a system which can be used by non-computer people familiar with the domain of interest and which incorporates the knowledge gleaned by human "experts" as a result of their experience in the area. The system uses this knowledge to exhibit expert behaviour by interacting with the user as an expert would, requesting information and using the resulting answers, together with volunteered information and its knowledge to draw the proper conclusions. The whole process is interactive, and high level interfaces involving graphical techniques and natural language processing are used to make the system friendlier to use.

The system is provided with tools so that "experts" can easily update the store of knowledge (the "knowledge base"). In particular, the knowledge is expressed in a declarative manner understandable to people familiar with the domain. The knowledge base and its support functions are usually a separate, identifiable component of the system.

Since the knowledge is expressed in domain-related terms and is likely "shallow" (in the sense outlined above), it is not immediately executable. This implies the need for a component in the system which can "interpret" the declaratively presented knowledge in some organized, directed way. This component is usually called the "inference engine".

Expert Systems and SDP

It seems unlikely that computer systems will replace "experts" in the short or medium term. This is because true experts in any area, no matter how focused, seem to require some general purpose common-sense reasoning capability and true artificial intelligence still seems pretty far away. In addition, computer systems (even expert systems) require maintenance and this implies ready access to experts who can maintain the knowledge base.

On the other hand expert systems seem quite likely to provide intelligent assistants to various areas of expertise. These are computer programs which behave like experts but require a bit of basic domain-specific knowledge for their successful use. The overall effect will then be to magnify the power of individual experts. Then experts can become even more specialized at the same time as their expertise becomes more distributed. Groups of experts will also have a focus for maintaining, sharing, and extending their expertise through an expert system.

Within the SDP framework, the following expert assistants could be envisaged:

- Data Analysis Assistant: This program could contain the expertise of an expert data analyst within a particular domain identified by subject matter and data source. Although a general purpose data analyst expert is probably too ambitious a project, an assistant with specialized knowledge of a particular area would know about particular anomalies which have caused problems in the past and must, as a result, be screened. It could advise on variance estimation techniques that are appropriate in the domain. Alternatively, as an adjunct of a general system, it could analyse the performance of the system, or suggest parameter values based on previous experience.

- Economist Assistant: This could provide advice to an economist on the relative qualities of various sources of data for a given purpose, as well as suggesting techniques for combining or analysing the data. Again this would have to be focused to a particular domain to be effective.

- Methodologist Assistant: This could be a repository for information about survey designs which have been used, available frames, and appropriate general systems to provide advice in the initial phases of survey design. It could also provide advice on the various aspects of data analysis before and after survey. This would have to be focused by, for example, frame or subject matter area in order to progress beyond vague recommendations.

- Subject Matter Statistician Assistant: The expert system for the subject matter statistician could advise on matters of survey control, and the relative potential for different approaches to collection and capture of data. This undoubtedly would be subject matter area specific.

- Clerical Assistant: Many of the cases handled by clerical processing are routine and can be handled quickly by direct mechanical approaches. Such a system would filter out these easy cases, freeing time for the more complex cases. The overall effect would probably be more uniformly processed data, together with a reduced level of tedium associated with routine cases.

- User Assistant: An expert user of Statistics Canada products knows what publications contain what information and what division to phone for specific requests for particular kinds of data. An expert user assistant would capture this information and at the same time make it available to users who are not so expert.

Another kind of "expert system" which might become quite useful is sometimes referred to as a "knowledge-based front-end" (KBFE). This type of software provides information about how to use a computer system, whether the system is appropriate to the given task, and gives advice on the setting of various parameters for good per-

formance. This type of system has been advocated for use with large software libraries where the choice of the appropriate subroutine requires a complicated decision procedure. In the context of SDP, a good candidate for a KBFE might be a particular general system that performs a well-defined function, and has a group of satisfied users who have accumulated experience with the pragmatics of running the system including the settings of various parameters. If, in addition, there are a number of potential users of the system who are handicapped by not having direct access to the accumulated body of experience but wish to use the system, it could very likely be feasible to use an expert system as a vehicle for the sharing of expertise.

For example, a knowledge-based front-end might be appropriate for record linkage. This has been a recurrent problem at Statistics Canada and as a result there is a large body of experience related to the theory, and application of these methods including a generalized system in active use (GIRLS). The methodologies lend themselves to an AI-based approach based on shallow approximate reasoning, as well as profiting from a data-analytical experimental approach. Furthermore, there is continuing interest in applying the technology to new areas resulting in a demand on the existing expertise. Other examples might involve an edit and imputation system or a system that advises about violations of confidentiality.

There is no claim that the list is exhaustive. It need only be recognized where expertise is available within the organization, where the domain is focused enough, and where there is an expected payoff from capturing the knowledge and making it more generally available.

Expert System Shells

After a number of expert systems were built, the implementers realized that if they removed the contents of the knowledge base they had a piece of software that could be used to build another expert system. This led to, for example, EMYCIN which was MYCIN with the knowledge about bacterial infections removed.

Thus an expert system shell should provide the software for maintaining the knowledge base, the inference engine for using the knowledge, and the high level interface. Of course, software developers of expert system shells now view their software in terms of its functionality and features and a number of expert system shells which did not originate with a particular expert system have appeared on the market. These provide a wide variety of features which are useful for expert system development but in a very real sense they end up being simply a programming environment for a very high level language. The knowledge is represented in a very high level declarative language using language specific editing tools (Knowledge DBMS). The inference engine interprets the language as it uses the knowledge to control a computation process in what in the final analysis must be a direct algorithmic manner. The interface tools are similar to those that should be provided with any good programming environment as part of library support.

The above view suggests that expert system shells should be evaluated on the programming environment that they provide as well as the appropriateness of the tool to the given task:

1. The language implied by the knowledge representation system and inference engine should be evaluated as other languages are, namely, for clarity of semantics, efficiency, etc. Of course, the language will be extremely very high level as appropriate for the intended applications.

2. The editing interface should be evaluated as other language specific editors are.

3. The interface tools should be high level, and easy to use, portable where possible, and where not provide full access to the capabilities of the machine with a clear demarcation between the two.

4. Software development tools such as debuggers, trace packages, and performance measurement tools should be provided. They should be easy to use and powerful.

ESS and SDP

Expert system shells may result in improved productivity for programmers especially those involved in the development of expert systems. Unfortunately many expert system shells have problems. There have been large systems with many bugs and unimplemented features, and have led application developers into difficulties. They

have traditionally only been available on specialized hardware and have produced stand-alone applications that cannot interface with other software systems.

These problems have been greatly alleviated but there remains the problem that without standards, expert system shells have evolved separately in many different directions resulting in the necessity to commit to a particular software vendor. Furthermore, the software is expensive and difficult to evaluate properly without the additional expense of installing a non-trivial prototype and comparing ease of development, ease of use of the delivered system, efficiency, and portability.

On the other hand, many knowledge-based systems have been developed using expert system shells and the market is forcing more competition on the part of vendors. This will undoubtedly lead to better systems available at more reasonable cost which are benchmarked, and perhaps even standardized.

Conclusions

There is good potential for the AI technologies to benefit statistical data processing, leading to better, easier-to-use, more adaptable general systems which can be developed more quickly, and at less cost.

Appendix A

EXAMPLES OF VERY HIGH LEVEL PROGRAMMING

To give a feeling of the three types of programming approaches let us consider the (artificial) problem of sorting a list.

In the *functional programming* approach one uses recursion to replace iteration and avoids the use of destructive assignment to variables. The following is an ML implementation of insertion sort for integers [14]:

```
fun sort nil = nil
  | sort(a::l) = insert(a,sort l);
fun insert(x:int,nil) = [x]
  | insert(x,l' as a::l)
          = if x <= a
            then x::l'
            else a::insert(x,l);
```

In words, the result of sorting an empty list (`nil`) is the empty list (`nil`). The result of sorting a list with first element `a` and remainder `l` is the result of inserting `a` into the result of sorting `l`. The result of inserting `x` into the empty list is the list `[x]`. The result of inserting `x` into a list `l'` which has first element `a` and remainder `l` depends on whether `x <= a` or not. If yes, then the result is formed by constructing a list with first element `x` and remainder `l'`. Otherwise, the result is obtained by inserting `x` into the shorter list `l` and forming a result with first element `a` and as remainder the result of inserting `x` into `l`.

Higher order functions are used to remove the dependencies on the builtin ordering relation.

```
fun sortgen cfn nil = nil
  | sortgen cfn (a::l)
          = insertgen cfn (a,sortgen cfn
l);
fun insertgen cfn (x,nil) = [x]
  | insertgen cfn (x,l' as a::l)
          = if cfn(x,a)
            then x::l'
            else a::insertgen cfn (x,l);
```

Here `sortgen` is a function which takes a single functional argument and returns a functional result. If the argument is an ordering function on the integers, and the result is a sort function on the integers that has exactly the same behaviour as the previous implementation:

```
fun lesseq(x:int,y) = x <= y;
fun sort list = sortgen lesseq list;
```

In the *logic programming* approach one observes that a list LS of elements is a sorted version of list L if and only if LS can be obtained from L as the result of a permutation transformation (a reordering) and LS is in order.

```
sort(L,LS) :-
     permutation(L,LS),
     ordered(LS).
```

Note that this is a high level specification of what sorting is all about and does not presuppose any particular algorithm. It is also a valid Prolog statement whose declarative reading is essentially that presented above. The expressions `sort( )`,

`permutation( )`, and `ordered( )` are called predicates. A predicate can be considered as a test on all possible pairs (L,LS) passing judgement on each pair as to whether LS is an ordered permutation of L.

The rest of the sort program is as follows:

```
permutation(Xs,[Z|Zs]) :-
    select(Z,Xs,Ys),
     permutation(Ys,Zs).
```

In words `[Z|Zs]`, a list with first element `Z` and remainder `Zs` is a permutation of `Xs` if `Z` is in the list `Xs` and when removed leaves `Ys`, and `Zs` is a permutation of `Ys`.

```
permutation([],[]).
```

The base case: the empty list (`[]`) is a (the only) permutation of itself.

```
select(X,[X|Xs],Xs).
```

In words, `X` is in a list of the form `[X|Xs]` and leaves the remainder `Xs`.

```
select(X,[Y|Ys],[Y|Zs]) :-
     select(X,Ys,Zs).
```

Alternatively, `X` is in a list `[Y|Ys]` leaving `[Y|Zs]` if `X` is in the list `Ys` leaving `Zs`.

```
ordered([X,Y|Ys]) :-
     less_eq(X,Y),
     ordered([Y|Ys]).
```

A list `[X,Y|Ys]` (first element is `X`, second element is `Y`, remainder is `Ys`) is in order if `X`is less than or equal to `Y` and the list `[Y|Ys]` is ordered.

```
ordered([X]).
ordered([]).
```

Two base cases: lists with one or zero elements are ordered.

Now the program just provided is a correct Prolog program, given of course that `less_eq/2` (`less_eq` with two arguments) has been defined. But how will it be executed?

Given a predicate to prove (a request to compute):

```
sort([2,4,1,3],Answer).
```

Prolog will systematically compute the permutations one by one of [2,4,1,3]:

```
[2,4,1,3]
[2,4,3,1]
[2,1,4,3]
[2,1,3,4]
[2,3,4,1]
[2,3,1,4]
[4,2,1,3]
[4,2,3,1]
[4,1,3,2]
...
```

Each of these will be tested against the "ordered" relation to determine whether the list is in order or not. The process will terminate after the permutation [1,2,3,4] is found.

This shows a strength and weakness of Prolog: it can execute what amounts to a specification though inefficiently. On the other hand Prolog can mimic the behaviour of a better sorting algorithm as the following implementation of insertion sort shows:

```
sort([],[]).
sort([A|L],Result) :-
    sort(L,LS),
     insert(A,LS,Result).

insert(X,[],[X]).
insert(X,[A|L],[X,A|L]) :-
    less_eq(X,A).
insert(X,[A|L],[A|Result]) :-
    not less_eq(X,A),
     insert(X,L,Result).
```

As an example of the *object-oriented* approach we will consider Smalltalk-80. In Smalltalk-80 there is a class of objects that are maintained in a sorted order called "SortedCollection". An instance of the object is created by sending a `sortBlock:` message to the class `SortedCollection` with, as argument, a block providing the ordering relation. The result of this is an empty object ready to receive values. As elements are sent to the an object from this class they are inserted into the correct place.

```
ls <- SortedCollection
    sortBlock: [:a :b | a<=b].
(((ls add: 2) add: 4) add: 1) add: 3
```

The code for the insertion sort itself would appear as a method associated with the class `SortedCollection`. To support sorting of some other class the convention is to provide a method `asSortedCollection:` which constructs and returns a sorted collection by this mechanism. Thus the array

with elements 2, 4, 1, 3 could be sorted by the following command:

```
#(2 4 1 3) asSortedCollection:
    [:a :b | a <= b]
```

In summary, functional programming provides an easy to use, and yet efficient language which is procedure oriented (without side-effects). Logic programming allows the programmer to write specifications or procedural descriptions and execute either. Object-oriented programming provides a procedure oriented description which is associated with the objects being operated on.

References

(1) Apple Computer, Inc. (1987). *Apple Human Interface Guidelines: The Apple Desktop Interface.* Addison-Wesley, Reading, Mass.

(2) Fellegi, I. P. and Sunter, A. B. (1969). *A Theory for Record Linkage.* Journal of the American Statistical Association, Vol 64, pp. 1183-1210.

(3) Genesereth, M. R. and Nilsson, N. J. (1987). *Logical Foundations of Artificial Intelligence.* Morgan-Kaufmann, Los Altos, CA.

(4) Goldberg, A. and Robson, D. (1983). *Smalltalk-80: The Language and its Implementation.* Addison-Wesley, Reading, Mass.

(5) Goldberg, A. (1984). *Smalltalk-80: The Interactive Programming Environment.* Addison-Wesley, Reading, Mass.

(6) Harmon, P. and King, D. (1985). *Expert Systems.* Wiley, New York.

(7) Lloyd, J. W. (1987). *Foundations of Logic Programming. (2nd ed.)* Springer.

(8) Parsaye, K. and Chignell, M. (1988). *Expert Systems for Experts.* Wiley, New York, N.Y.

(9) Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ.

(10) Statistics Canada (1987). *Census of Canada 1986: Population and Dwelling Characteristics: Age, Sex and Marital Status* Statistics Canada Publication No. 93-101.

(11) Steele, G. L. (1984). *Common LISP.* Digital Press.

(12) Sterling, L. and Shapiro, E. (1986). *The Art of Prolog.* The MIT Press, Cambridge, Mass.

(13) Taylor, W. A. (1988). *What Every Engineer Should now About Artificial Intelligence.* The MIT Press, Cambridge, Mass.

(14) Wikstrom, A. (1987). *Functional Programming using ML.* Prentice-Hall.

A TECHNOLOGICAL FORECAST OF THE ROLE OF PERSONAL COMPUTERS IN STATISTICAL SURVEY METHODS

J. C. NASH,

University of Ottawa

A technological forecast of the role of personal computers in statistical survey methods.

John C. Nash
Faculty of Administration
University of Ottawa
Ottawa, Ontario, K1N 6N5

Abstract

Technological forecasts -- predictions of the methods and tools which will be used in the future -- are notorious for their lack of precision. Nevertheless, they often assist planners and managers by pointing out directions where development may occur or by suggesting likely opportunities or dangers. This presentation will attempt a technological forecast for personal computers and work-stations as they may be used to carry out and analyze statistical surveys. In particular, personal computer issues for discussion will include communications capabilities, data manipulation software, storage technology, and user education. These will be related to the tasks of planning, conducting, analyzing and reporting of statistical surveys.

Introduction

Technological forecasts are predictions of the methods, tools, and more generally situations which will pertain in the future. They are notorious for their lack of precision and for the glaring errors, both of commission and omission, which attend their publication. In this paper, I will nevertheless attempt to provide some technological forecasts relating to personal workstations and other "high technology" which may be likely to appear for carrying out (statistical) surveys.

At the outset, it is pertinent to note the major failings of technological forecasts. First, they are almost always wrong in some aspect. Both the timing and the nature of changes are notoriously difficult to gauge. Unanticipated developments -- in technology, marketing, social change, or government action -- can wildly alter the background in which technological developments take place.

Nevertheless, forecasts often assist planners and managers by pointing out directions where development may occur or by suggesting likely opportunities or dangers. Indeed, in the agricultural sector, forecasts of shortage or glut are sometimes made and communicated to producers with the express intent of permitting adjustments in output. The adjustments may make the forecasts "wrong", but the forecast may be considered to have been useful.

The presentation which is a companion to this paper is scheduled at the end of Symposium 88. As such, the oral presentation will naturally mention ideas discussed by other speakers which have not been considered here.

What are some of the methods of technological forecasting? The most

simple is the "seat of the pants" opinion of someone considered to be expert in an appropriate field. The extension of this to a panel of experts is generally favoured over one person's opinion, and a prominent Canadian example is the annual Agricultural Outlook Conference. Panels may, however, suffer from one or more strong personalities influencing the other contributors. The Delphi method seeks to overcome such psychological factors by isolating the panelists. To allow for revision of opinions, a referee circulates the opinions and amendments one or more times to the panelists, and reports the consensus (or lack thereof) at the end of the exercise.

These methods are essentially qualitative in nature. To them may be added Scenario Writing, where an individual or team tries to suggest possible directions of events. This may be a purely qualitative exercise, but some workers use dice or computer generated pseudo-random choices to select directions for development of the scenario (Sutherland, 1975). Nevertheless, the success of scenario techniques relies heavily on the skill of its users.

A more quantitative family of methods I will term "extrapolations". As technology advances, we see various measures of performance or cost changing with time. Fitting a curve to the plot of the measure versus time allows an extrapolation to be made. Examples of such plots and their extrapolations include cost per kilobit of memory, transistor gates per chip, or cost per megabyte of disk storage. The major failing of such extrapolations is that they may miss the introduction of a totally new form of technology, or may ignore fundamental limits to technological development, although knowledgeable users generally consider the latter possibility.

Growth curves are a special case of extrapolations, having particular utility in predicting the adoption of new technologies. In particular, the sigmoid curve (logistic, Gompertz, or similar) can be used to model the proportion of a population which has changed to a new technology. I have personal experience with this approach in the field of dairy production (Nash and Teeter, 1975; Nash, 1977, Nash and Walker-Smith, 1987). Similar techniques exist for many other applications (for example, Fisher and Fry, 1983). A weakness of the method is that we generally need to have rather more data than is available in order to predict the location of the asymptote and the timing of arrival within some distance of this asymptote.

To overcome the lack of data, forecasters may substitute into the growth model parameters drawn from models of similar technological changes which are farther along the change process. The troublesome word here is "similar", along with the possibly heroic assumption that the rate of adoption of "new" ideas will follow the same pattern in two different situations.

A parallel idea is to use one measure (variable) as a predictor of another -- the concept of leading indicators (Makridakis et al., 1983, Chapter 12). While these methods are much beloved by stock analysts,

my own opinion is that the measures chosen as indicators are either obvious, thereby giving very little information about future events that is not common knowledge, or work over such a limited range of events or time intervals to be of little practical utility.

As may be obvious, all the quantitative methods mentioned so far are applicable primarily in the short term. For the longer term -- by which I mean greater than 3 to 5 years -- there are few methods. Long-term structural models, primarily of the econometric variety, offer one possibility. However, they rarely account for technological change. Indeed, they employ a more or less explicit assumption that the processes and trends of past experience will carry forward into the future.

A method which may allow the forecaster to provide some quantitative information to long-term technological development is cross-impact modelling (see Turoff, 1975, as well as other references below). This can be both a conceptual and a calculation tool. The principal idea is that the development of one technology either enhances or diminishes the likelihood or rate of development of another. For example, the development of electronic cash registers halted development of electromechanical devices.

The main weaknesses of cross-impact analysis are that the user may fail to include a relevant technology or may provide poor estimates of the relevant weightings. Nevertheless, this analysis does improve on "guessing" about the future, and is the main form of analysis which will be used here. However, I consider this approach, in its present form as used here at least, to be primarily a tool for structuring my thoughts. In preparing this paper, the development of a program to extend the ideas which will be discussed was attempted, but I feel there are so many choices and assumptions made in the implementation of this program that time and space restriction preclude further mention of it here.

This presentation will attempt a technological forecast for the impact of personal computers and work-stations as they may be used to carry out and analyze statistical surveys. The time horizon for the forecasts is the year 2000, that is 12 years hence.

Survey activities

Table 1 presents a view of the work-station technologies involved in various aspects of conducting a survey. The types of software which could be employed are also listed. I believe that examples exist now of all the types of software mentioned. However, my opinion is that no consolidated and complete system has yet been implemented, even though the technological basis does exist. In part, this may be due to conflicts or incompatibilities in the data structures, standards, and computing environments used in different sub-systems. More likely, the legacy of existing organizational, data processing, and personnel

resources, together with budget restraints, hamper the development of a unified work-station statistical survey system.

For surveys which do not address questions of human respondents, we may see very interesting changes in survey methodology. Consider, for example, the interrogation of instruments by remote communications. This can be and is already done using standard modem technology with an auto-answer capability. Remote utility meter reading is already under consideration. Normally such applications would be used periodically for billing purposes, but the use in a sample survey to determine the nature of demand for the utility's services can be considered once the appropriate equipment is in place.

Such an example underlines several of the issues addressed later in the paper:

1) users may feel the survey use of such equipment to be an invasion of their privacy

2) investment costs are such that not all users may be supplied with the "new" equipment

3) employees are replaced with equipment and join the ranks of the unemployed.

Despite potentially untoward uses of automated surveys of this kind, I believe that in the next 15 years such applications will grow. Particular areas of application include environmental monitoring (both indoor and outdoor), quality control and improvement of manufacturing and services, human and vehicular traffic movement, wildlife or livestock management, financial auditing and similar surveys of computerized book-keeping or point-of-sale records, and marketing and advertising surveys based on noting television channel activity.

Existing Technological Flows

Extrapolation of existing flows in the development of work-stations and related technologies is relatively straightforward. A good example of such an extrapolation is Tablet, a lap-top portable computer (Mel et al., June 1988). This is a design for the year 2000. However, the cover of the August 1, 1988, Fortune Magazine shows John Scully, President of Apple Computer, holding the Knowledge Navigator prototype, which looks and is briefly described in such a manner that it appears to be an implementation of Tablet.

One of the important suggestions for Tablet is the use of a screen display on which the user writes with a stylus. Handwriting recognition software designed to "learn" the user's script and translates it in real time. Thus the user is able to correct errors in the translation, reducing the need for the software to have an extremely low error rate. Tablet has obvious applications at the

interview stage of survey taking, but offers many other possibilities elsewhere in the process.

More conventional extrapolations concern the areas of communications, data manipulation, storage technology, and user education.

Communications facilities for data transmission are improving steadily. Despite the security and confidentiality obstacles, it seems likely that statistical surveys will increasingly use data rather than voice communications in data gathering. Other important possibilities relate to teamwork in the analysis of survey data by workers physically remote from each other and to dissemination of the results of the survey to clients.

Data manipulation SHOULD improve with the increasing sophistication of software. However, commercial interest may hinder the easy movement of data. An example concerns the inability of some versions of Lotus 1-2-3 to read worksheet files created by other versions of the same package. The motivation for this incompatibility seems to be to sell "updated" versions of the software, but the issue is clouded by the presence of "enhancements" in the newer or non-student versions. Other packages have similar version conflicts.

Devices for data storage have shown a steady growth in capacity per dollar. This is easily documented for both flexible and fixed magnetic disk devices. Optical data storage has been heralded for a number of years, but its market diffusion and price changes appear to be much less rapid than anticipated in early press releases and commentary. Nevertheless, devices are available which store 200-1000 megabytes of data, allowing workstations to access the returns from very large surveys and to store graphic images representing, for example, the streets of a municipality where a survey is to be conducted.

User education is one area where the promise of personal computers has yet to be widely realized. Computer aided instruction software ("courseware") has proven very expensive to prepare and install. (At the Statistical Society of Canada 1988 Annual Meeting in Victoria, Dr. J.C. Holt of the University of Guelph was one of several participants who described efforts to develop such courseware.) It would seem that CAI for various aspects of survey sampling would be a highly effective tool. Potential consumers of such training, perhaps focussing on a sub-topic of statistical surveys, are often geographically scattered and/or have schedules which do not permit them to congregate at a single time and place for training. Because of the costs of developing CAI courseware, the social aspects of education, and the fact that it may tend to put the developer out of a job, I do not forsee rapid developments in this area. Indeed, I suspect that by 2000 we will see CAI take a larger, but still very minor, role in education generally, and have a place in statistical survey training only where special efforts are taken to establish it.

Impact of technology on survey taking

It is my opinion that the direct impact of advances in workstations and related technologies are moderately straightforward to forecast. The steady, possibly inexorable, march of progress in speed and capacity of workstations will naturally make it possible for a very small statistical team to carry out quite large surveys which do not require personal interviews (i.e. those done by mail or electronic communications or which do not otherwise require human labour). In the previous section, a more detailed technological assessment has already been presented. The questions which remain would seem to be those of the integration of people, software, hardware and operating procedures into systems which operate smoothly. That is, I perceive the major obstacles to be those of incompatibilities in the requirements and capabilities of surveys and the means to execute them.

Less easy to predict are the more generalized impacts of these developments. However, with the temerity of one whose livelihood is not directly drawn from survey taking, I will venture some forecasts, prefacing them with the statement that I have not yet been able to provide quantitative support for my suggestions.

First, it seems likely that the overall cost of conducting a survey through all stages from planning to reporting should decrease with the reduced labour workstation technology allows. Pressures in governments to cut the size of civil services may find an outlet through application of this technology. A counter pressure exists, however, in the trend toward linking government programs, including transfer payments of various kinds, to the results of surveys. In Canada, cost of production surveys are mandated to set prices for several food products, as just one example of this tendency. Informal application of survey results is also vigorously pursued by elected representatives and interest groups to demand action to address perceived disadvantage in employment, income, education, medical services and so on. The statistics are clearly important.

Second, because of this importance, we may expect efforts to be taken to cause survey results or interpretations to support a particular point of view. With the widespread availability of electronic technology and expertise in its construction and operation, it seems likely that corruption of surveys for economic or political gain will become a difficulty for statisticians to contend with. I do not suggest that such sabotage will be pervasive, but warn that the concentration of the control of surveys into fewer human hands with more operations taking place electronically makes it easier for a survey to be corrupted. Even without any interference with survey conduct or data, the technology makes it easier and cheaper to 'train' respondents. Several years ago in Canada a special interest group held a seminar on how a survey questionnaire should be answered to provide results most favourable to the group. Besides such falsification for advantage, we may expect destruction of information by "hackers" -- computer vandals -- or those with some interest in preventing the

recording of information.

If information is destroyed or corrupted, I suspect a large segment of the general population would, rather than expressing outrage, actually applaud the culprit(s). Privacy concerns are increasing, and "human rights" are a growth industry (Dyer, 1988). The successful legal challenge of the Canadian Census questionnaire (see references below), whatever its legal or moral merits, serves to underline this trend. The use of linkage of files by government or private organizations to "snoop" on citizens is immensely aided by the development of artificial intelligence software and new hardware tools such as the Connection Machine (Stanfill and Kahle, 1986). Similarly, the capacity of agencies to survey without formal questionnaire or even the knowledge of the respondent is a possibility the general population will not accept easily. In the public mind, the ability of large organizations to properly handle data which relates to individuals is generally held in low regard.

Despite concerns for individual "rights" (or at least privileges), our society appears to be splitting into two major segments: those who belong by owning property and having income and those who survive on the fringes -- the homeless, the handicapped, the criminal, the addicted and the insane. I refer to the fringe society in Table 2 as dispossession for want of a general term. Technological advances may increase this schism in society by rendering more individuals incapable of earning a living as their jobs are taken over by machines or de-skilled to the extent that they pay too little compared to welfare or crime. For the statistician trying to conduct a survey, the dispossessed presend a formidable challenge. They cannot easily be situated in a sampling frame, may react with violence to attempts to interview them, and may respond with whatever answers appear to them most likely to be personally rewarding. Present government policies in the major Western countries do not appear directed to halting the trend to a greater percentage of the population on the fringe. The bibliography points to several references, two of which imply a survey of homelessness has been carried out, but at the time of writing I have not been able to determine the methodology of this survey.

In summary, statisticians should expect to have an easier time planning, analyzing and reporting surveys, and a much tougher job actually getting hold of the right data. When quality data is collected, however, the tools should allow for better information to be extracted from surveys, with increased timeliness of reporting, precision and reliability of estimates, and finer grain of analysis in terms of time periods, geographic areas or other categorizations. Furthermore, the reporting can and will include better graphical aids and should avoid typographical errors by having tables directly typeset into the report from the most current analysis.

Conclusion

As this forecast is published in the proceedings of this conference I am in the awkward situation that my words may return to haunt a future meeting. To the extent that some of the above discussion is a warning of what, as statisticians, we may wish to work to avoid or prevent, I sincerely hope that some of my predictions are not realized.

Acknowledgements

Research by the author is partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. Telex communications to arrange this contribution were provided by the Department of Scientific and Industrial Research of New Zealand, Mount Albert Research Centre, Auckland. Conversations with Kevin Price, Gordon Sande, Elaine Hoskins, and Mary Nash were helpful in developing some of the ideas presented.

References

Alberta (Western) Report, "Beating Back Census", v. 14, n. 48, Nov. 16, 1987, p. 42 and v. 15, n. 17, April 11, 1988, p. 24.

Calgary Herald, "Businessman wins fight over census, March 23, 1988, p. B6.

Cetron, M.J. and Ralph, C.A., Industrial Applications of Technological Forecasting: Its Utilization and Management, Malabar, Florida: Robert E. Krieger Publishing Co., 1983.

Dalby, James F., "Practical refinements to the cross-impact matrix technique of technological forecasting", contribution XV in Cetron and Ralph (1983).

Dyer, Gwynne, "Human rights -- and who is human", Ottawa Citizen, Tuesday August 16, 1988, p A8.

Fisher, J.C. and Fry, R.H., "A simple substitution model of technological change", contribution XVII in Cetron and Ralph (1983).

Fortune Magazine, August 1, 1988 issue, Cover photo.

Halifax Chronicle Herald, "Halifax homeless on increase", Nov. 29, 1986, p. 1,2.

Jantsch, Eric, Technological Forecasting in Perspective, Paris: OECD, 1967.

Linstone, H.A. and Turoff, Murray, The Delphi Method: Techniques and Applications, Reading, MA: Addison-Wesley, 1975.

Makridakis, S., Wheelwright, S.C., and McGee, V.E., Forecasting: Techniques and Applications, 2nd Edition, New York: Wiley, 1983.

Mel, B.W. et al. , "Tablet: personal computer in the year 2000" Comm. ACM, v. 31, n. 6, PP 638-646, June 1988

Montreal Gazette, "Census violates privacy, commission rules", June 21, 1988, p. A7.

Nash, J.C. "A discrete alternative to the logistic growth function", Applied Statistics, v. 26, n. 1, 1977.

Nash, J.C., and Teeter, N.J., "Building models: an example from the Canadian dairy industry", Canadian Farm Economics, v. 10, n. 2, pp. 17-24, April 1975.

Nash, J.C. and Walker-Smith, M., Nonlinear Parameter Estimation: an Integrated System in BASIC, Marcel Dekker Inc.: New York, 1987.

Stanfill, C. and Kahle, B., "Parallel free-text search on the Connection Machine system", Communications of the ACM, v. 29, n. 12, December 1986, pp. 1229-1239.

Sutherland, J.W., "Architecting the future: a Delphi-based paradigm for normative system-building", contribution VI.E. in Linstone and Turoff (1975).

Toronto Globe and Mail, "Reagan's cutbacks blamed as US homeless toll grows", Feb. 6, 1987, p. A10.

Toronto Star, "Group plans survey of homeless", Jan. 14, 1987, p. A3.

Toronto Star, "100,000 Canadians homeless during 1986", April 7, 1987, p. A8.

Toronto Star, "Shelters forced to turn away homeless", April 8, 1988, p. A1, A4.

Toronto Star, "200,000 families said to lack decent housing", June 28, 1988, p. A3.

Turoff, Murray, "An alternative approach to cross impact analysis", contribution V.C. in Linstone and Turoff (1975).

Winnipeg Free Press, "Cutbacks increase (Edmonton's) homeless", August 10, 1987, p. 40.

Winnipeg Free Press, "250,000 Canadians homeless -- national survey", Oct. 23, 1987, p. 10.

Winnipeg Free Press, "8 Manitobans in hot water over census", May 3, 1987, p. 4.

Table 1: Work station technological needs for aspects of statistical surveys.

Aspects of surveys of interest:	computer tools
- planning	
- initiation of project: needs, goals	Text processing
- planning of survey instrument	Text processing
- sampling design	Calculation, simulation
- frame development	Database, storage
- costing	spreadsheet
- data collection	
- "interview" or equivalent	specialized software
- imputation	database
- monitoring of process (Quality Improvement)	project management tools, specialized software
- training of staff	CAI
- simulations	simulation software + specialized enhancements
- management	project management tools
- data analysis	
- aggregation	database
- edit	database
- imputation	database (specialized)
- outlier detection and display	graphics/calculation/database
- process monitoring and log development	specialized or general project management tools
- estimation processing	various
- report preparation	
- text	text processing --|
- graphics	graphical tools |- integration
- tables	spreadsheet --|
- interpretation	expert systems

Table 2. Technologies or situations considered

1. Computer improvement -- More powerful hardware and software with better user interfaces

2. Dispossession -- homelessness, unemployability, illiteracy, de-skilling of jobs, etc.

3. Human rights and privileges -- growth in privacy, personal security, individual freedoms, anti-discrimination and environmental concerns

4. Number dependency -- legal requirement or other need for the results of statistical surveys

5. Communications enhancements -- easier transfer of information, compatibility of data systems, standards, etc.

Table 3. Cross-impact of different "technologies" on each other on a scale -2 to 2 (--, -, 0, +, ++)

Scale of effect of technology at left on technology at top of table is given in the entry in the body of the table.

	Computer improvement	Dispossession	Human Rights	Number Dependency	Communications
Computer improvement	++	+	-	++	++
Dispossession	0	+	+	--b	-b
Human Rights	--	-a	+	-	-
Number Dependency	+	0	-	+	+
Communications	+	+	0c	+	+

a -- this element is negative, since improved concerns about human rights and privileges will REDUCE dispossession.

b -- inability to place people in a sampling frame may make it very difficult to include them in a survey. However, a view may be taken that those who cannot afford any of the products of society do not need to be considered. Here, I have taken the view that all citizens are part of the society.

c -- communications technology may have profoundly negative impacts on rights and privileges, but at the same time may be used positively. The author has personal knowledge of several instances where serious human rights violations were avoided or minimized by Amnesty International members and supporters using communications technology intelligently.

Statistics Canada
Statistique Canada

Ottawa, Canada

SYMPOSIUM 88

Les répercussions de la technologie de pointe sur les enquêtes

24 et 25 octobre 1988

Ottawa (Ontario)

ACTES

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010438786

Canada

SYMPOSIUM 88

Les répercussions de la technologie de pointe sur les enquêtes

24 et 25 octobre 1988

Ottawa (Ontario)

ACTES

Publiés en avril 1989

Compilés et révisés par

J. Kovar et E. Doucet

PRÉFACE

Le Symposium 88 fut le cinquième symposium international parrainé par Statistique Canada sur les questions méthodologiques et informatiques. Comme par le passé, l'objectif du symposium visait à réunir des experts du monde entier pouvant échanger des idées d'intérêts communs.

La conférence a eu lieu au Centre de conférences Simon A. Goldberg de Statistique Canada. Plus de 400 personnes, représentant plus de 20 pays, y ont participé. Les allocutions d'ouverture ont été suivies du discours-programme de Bo Sundgren intitulé "L'impact de la nouvelle technologie sur la méthodologie et l'organisation du traitement des données statistiques". Dix-sept communications sollicitées ont ensuite été présentées au cours des deux jours prévus à cette fin. L'emphase a été mise sur les utilisations novatrices de la technologie de pointe relativement à la collecte, le traitement, l'analyse et la diffusion des résultats.

Ce volume des Actes du Symposium contient un registre des communications telles qu'elles ont été soumises. Les vues et les opinions exprimées ici sont celles des auteurs et ne représentent pas nécessairement les vues ou politiques officielles des organismes pour lesquels ils travaillent, pas plus que celles de Statistique Canada. Dans le but de faciliter la consultation de ces Actes, la table des matières de ce volume correspond à l'ordre de la présentation du programme.

Le sixième Symposium international de Statistique Canada se tiendra à Ottawa, en octobre 1989. Il portera sur l'analyse des données dans le temps, l'emphase étant particulièrement mise sur le développement et l'analyse des enquêtes périodiques. Nous espérons vous y retrouver et profitons de cette occasion pour remercier tous ceux et celles qui ont contribué à faire du Symposium 88 un succès.

J.E.Doucet
Direction de l'informatique
Statistique Canada

J.G.Kovar
Direction de la méthodologie
Statistique Canada

TABLE DES MATIÈRES

Allocutions d'ouverture: I.P.Fellegi, Statisticien en Chef du Canada .. 1
(Présenté par **G.Brackstone,** Statistique Canada)

Discours-programme: L'impact de la nouvelle technologie sur la méthodologie et l'organisation du traitement des données statistiques . 7
B.Sundgren, Statistics Sweden

Session 1: Derniers développements
Présidé par **E.Outrata**, Statistique Canada

L'utilisation de nouvelles technologies à Statistique Canada 31
J.E.Doucet, Statistique Canada

L'utilisation de la technologie à Statistics Sweden 45
L.Lyberg et P.Dean, Statistics Sweden

HyperQuest: Une nouvelle méthode pour interviewer à l'aide des ordinateurs portatifs 69
M.A.Jaro, Systems Automation Corporation

Session 2: La collecte des données
Présidé par **B.Petrie,** Statistique Canada

L'impact de la technologie de pointe sur la collecte des données ... 79
W.L.Nicholls II, US Bureau of the Census

Les effets de l'ITAO sur les coûts et la qualité des données 113
G.Catlin, S.Ingram et **L.Hunter,** Statistique Canada

Deux nouvelles technologies de collecte expérimentées à l'INSEE 131
C.Bernard, Institut national de la statistique
et des études économiques

Les banques informatisées de questions: Pourquoi et comment? 147
R.P.Baker, National Opinion Research Center

La collecte et la saisie des données pour les enquêtes 165
I.S.Green et M.J.Wade, Statistique Canada

Session 3: Le traitement des données
Présidé par **J.Ryten,** Statistique Canada

Le traitement d'enquête assisté par ordinateur sur le PC et le PC-LAN 183
W.J.Keller et J.G.Bethlehem,
Netherlands Central Bureau of Statistics

L'utilisation des micro-ordinateurs pour le traitement des données de recensement dans des pays en voie de développement: Un mise à jour 197
V.Toro et K.Chamberlain, US Bureau of the Census

Un rapport sur le traitement d'enquête réussi dans des pays en voie de développement: L'approche des micro-ordinateurs au niveau des enquêtes démographiques et de santé 215
J.Cushing, Institute for Resource Development/Westinghouse

Une comparaison de plusieurs logiciels pour vérifier les données d'enquête et de recensement 227
P.Cotton, Consultant en Statistique

Session 4: L'analyse des données
Présidé par **J.N.K.Rao,** Carleton University

Logiciel pour l'analyse d'enquête 259
W.A.Fuller, Iowa State University

SUDAAN: Un progiciel détaillé pour l'analyse des données d'enquête .. 267
L.M.LaVange, **B.V.Shah,** B.G.Barnwell et J.F.Killinger,
Research Triangle Institute

Session 5: Une vue sur l'avenir
Présidé par **G.Brackstone,** Statistique Canada

Contributions de l'intelligence artificielle au traitement des données statistiques 285
J.H.Johnson, Statistique Canada

Un prévision technologique du rôle des ordinateurs personnels au niveau des méthodes statistiques d'enquêtes 305
J.C.Nash, l'Université d'Ottawa

ALLOCUTIONS D'OUVERTURE

I. P. FELLEGI,

Statisticien en Chef du Canada

Présenté par G. Brackstone,
Statistique Canada

BIENVENUE

Bonjour. Je vous souhaite cordialement la bienvenue au Symposium de 1988. Il s'agit du cinquième de la série de symposiums parrainés par Statistique Canada, le Laboratoire de recherche en statistique et probabilité de l'Université Carleton et l'Université d'Ottawa.

Cette année, les organisateurs nous ont préparé un programme imposant. Il m'a fait particulièrement plaisir de constater qu'à l'échelle internationale, la participation était bonne. Sont bien représentés:

- Statistique Suède (Statistics Sweden),
- L'Institut national de la statistique et des études économiques de la France,
- le Bureau de la statistique des Pays-Bas (the Netherlands Bureau of Statistics), et
- le U.S. Bureau of the Census;

de même que plusieurs universités et sociétés privées bien connues au Canada et aux États-Unis.

Au nom de Statistique Canada et de la ville d'Ottawa, permettez-moi de vous souhaiter chaleureusement la BIENVENUE et WELCOME!

Bienvenue aussi au grand nombre de délégués de Statistique Canada. Je vois que nous sommes bien représentés: à la fois au programme officiel et dans l'auditoire.

COMMENTAIRES SUR LE CHOIX DU SUJET

Les précédents symposiums de cette série portaient surtout sur des sujets directement reliés aux techniques d'enquête. Au fil des ans, nous nous sommes réunis dans des circonstances semblables pour traiter des questions suivantes:

- L'analyse des données d'enquête (1984);
- La statistique des petites régions (1985);
- Les données manquantes dans les enquêtes (1986);
- Les utilisations statistiques des données administratives (1987).

Le sujet de cette année, LES RÉPERCUSSIONS DE LA TECHNOLOGIE DE POINTE SUR LES ENQUÊTES, représente un léger écart par rapport à la tendance qui s'est dessinée au cours des quatre dernières années. Ce sujet se situe clairement à la croisée des techniques d'enquête et de l'informatique, et je considère qu'il est très opportun d'en discuter. Pourquoi? Parce que ce sujet présente un défi à la fois au MÉTHODOLOGISTE D'ENQUÊTE et au SPÉCIALISTE DE L'INFORMATIQUE. Permettez-moi, pendant quelques minutes, de m'interroger sur la nature de ce défi.

Défi lancé aux informaticiens

Examinons d'abord la nature du défi que doit relever le spécialiste de l'informatique. Au cours des dernières années, Statistique Canada a connu une période de croissance et d'innovation sans précédent dans le secteur de l'informatique. Nous n'en manquons pas d'exemples.

- Nous disposons maintenant, dans cet organisme, de 1200 postes de travail articulés sur un micro-ordinateur; il y a quatre ans nous n'en avions aucun.

- À Statistique Canada, nous possédons une unité centrale moderne dont la puissance s'est accrue régulièrement à un rythme d'environ 30 % par an au cours des dix dernières années.

- Nous avons accès au plus récent logiciel de base de données relationnelles par différents genres de systèmes d'exploitation, ce qui permet d'élaborer très rapidement des applications assez sophistiquées.

- Nous avons aussi commencé à profiter de technologies qui, jusqu'à tout récemment, étaient perçues comme ésotériques: postes de travail d'infographie interactive, mémoires à disques optiques, systèmes d'éditique, etc.

Cependant, même si nous sommes entourés des signes tangibles du progrès accompli, beaucoup des problèmes contre lesquels nous avons lutté pendant une quinzaine d'années ne sont toujours pas résolus. Considérons seulement quelques-uns de ces problèmes persistants.

- PROBLÈME - Nous n'avons toujours pas trouvé de méthode automatisée qui, tout en étant efficace et abordable, nous permettrait d'éviter de divulguer les renseignements confidentiels.

- PROBLÈME - L'impressionnante puissance de calcul dont nous disposons est encore insuffisante pour répondre aux besoins de conception et de spécification des enquêtes. De plus, beaucoup d'analystes ne sont que trop disposés à construire un autre système de correction des données pour une autre enquête, satisfaits de réinventer la roue et de faire quelques petits ajustements marginaux.

- PROBLÈME - Malgré le fait qu'un très puissant logiciel privé soit disponible, l'extraction et le rassemblement des données sont encore sérieusement limités parce qu'il est difficile de se retrouver dans le dédale des technologies qui servent à stocker les données recherchées.

Il ne s'agit là que de quelques-uns de ces problèmes persistants; vous avez probablement votre propre liste. Le Symposium de 1988 met le spécialiste de l'informatique au défi d'explorer "de nouveau" les possibilités d'appliquer les fruits de cette innovation technologique au bénéfice plus direct du processus d'enquête.

Défi lancé au méthodologiste

Le défi lancé au méthodologiste d'enquête par ce symposium est peut-être moins évident.

Le fait que de puissantes ressources informatiques soient disponibles à relativement peu de frais a ouvert la prote à une réévaluation de notre choix de méthodes et de techniques d'enquêtes. Permettez-moi de vous montrer ce que je veux dire à l'aide de quelques exemples:

- Les techniques statistiques qui exigent beaucoup de ressources informatiques sont devenues de plus en plus réalisables (du point de vue économique et technique) et il existe un grand nombre de façons intéressantes de les appliquer; échantillonnage en plusieurs étapes, simulation et modélisation, analyse de données complexes, approche exhaustive du couplage de fichiers, pour n'en nommer que quelques-unes. Nous devons nous demander régulièrement si la façon dont nous abordons les problèmes que nous voulons résoudre est appropriée, compte tenu de ces nouvelles possibilités, ou si, professionnellement, nous nous plions à la "pratique généralisée" et aux "convenances" parce que nous avons l'habitude des techniques que nous utilisons et que, même, nous nous sentons à l'aise.

- Le traitement d'enquêtes axé sur les bases de données est devenu réellement possible avec l'avènement de logiciels de systèmes de gestion de bases de données relationnelles qui sont commercialement viables. L'approche axée sur les bases de données offre la possibilité de rompre avec le modèle traditionnel "de traitement chronologique en lot" du traitement d'enquêtes, modèle qui reflète les contraintes technologiques imposées par les précédentes étapes du développement de l'informatique. Le modèle que nous utilisons présentement pour le traitement d'enquêtes semble être de ce genre-là, parce qu'il a dû être transformé pour pouvoir fonctionner avec les précédentes générations d'ordinateurs. Maintenant que les unités de traitement de bases de données relationnelles sont disponibles, il est peut-être temps de "repenser" notre modèle de traitement d'enquêtes.

Le symposium de 1988 met le méthodologiste d'enquête au défi d'examiner le répertoire actuel de modèles, d'outils et de techniques, et d'en évaluer, aujourd'hui, la pertinence, compte tenu du vaste éventail de possibilités qu'offrent les plus récentes technologies. En outre, il est encore plus important de savoir si le fait d'aborder nos problèmes de façon non traditionnelle nous conduira aux solutions qui nous ont échappé jusqu'ici.

Synergie

Bien entendu, les défis que doivent relever l'informatique et la méthodologie ne peuvent en réalité pas être aussi facilement compartimentés qu'il est possible de le faire dans un discours inaugural comme celui-ci. De plus en plus, il est vrai que:

- le méthodologiste d'enquête connaît très bien l'informatique et

- le spécialiste de l'informatique a une bonne connaissance de la statistique et des techniques d'enquêtes.

Voilà pourquoi nous pouvons nous rencontrer, dans le cadre du Symposium de 1988, pour relever ensemble ces nouveaux défis.

Cela nous ramène à la croisée de l'informatique et de la méthodologie, et à la synergie dont les mécanismes, comme ceux du Symposium 1988, peuvent permettre de faire évoluer notre science. Le sujet de notre rencontre a été bien choisi. Il est opportun et le restera pour les années à venir.

PRÉSENTATION DU CONFÉRENCIER INVITÉ

Je passe maintenant à la tâche agréable de présenter le conférencier invité du Symposium de 1988, qui est aussi notre collègue, Monsieur Bo Sundgren, qui nous vient de la Suède.

Bo Sundgren est:

- Chef du département des systèmes de traitement électronique des données à Statistique Suède,
- maître de conférence à l'Université de Stockholm, et
- maître de conférence à la Stockholm School of Economics.

L'association entre Statistique Canada et Bo Sundgren dure depuis longtemps et a été profitable. Ceux d'entre vous dont l'expérience du traitement des données statistiques remonte au début des années soixante se souviendront très bien de Bo Sundgren. C'est un précurseur important dans les domaines connexes de la modélisation des données et de la conception de bases de données, un précurseur qui connaissait bien les besoins spéciaux d'un bureau statistique. En particulier, il a joué un grand rôle dans les débuts de l'élaboration d'une unité de traitement relationnelle pour les systèmes de recensement, à une époque où aucune n'était disponible sur le marché. Ses efforts ont contribué au développement du système RAPID présentement utilisé dans de nombreux pays (le nôtre inclus).

Au fils des ans, nous avons continué de collaborer avec Bo Sundgren, principalement dans le cadre de son travail avec le Programme de calcul statistique des Nations Unies et de la Conférence des statisticiens européens. Dans ce domaine, on reconnaît généralement qu'il guide les efforts faits internationalement pour uniformiser les logiciels de traitement d'enquêtes qui pourraient être utilisés par les principaux bureaux statistiques nationaux et servir de base à un transfert de technologie vers les pays moins développés.

Bo Sundgren est un grand penseur qui n'a jamais hésité à examiner de plus près les principes de base qui sous-tendent souvent notre façon d'aborder un problème. Je suis convaincu qu'il suscitera en vous la réflexion et qu'il donnera le ton à ce qui sera, je l'espère, un symposium enrichissant et agréable.

Monsieur Sundgren, bienvenue au Symposium de 1988.

DISCOURS-PROGRAMME: L'IMPACT DE LA NOUVELLE TECHNOLOGIE SUR LA MÉTHODOLOGIE ET L'ORGANIZATION DU TRAITMENT DES DONNEÉS STATISTIQUES

B. SUNDGREN,

Statistics Sweden

ABRÉGÉ

Au cours des dernières décennies, nous avons été témoins d'importants progrès technologiques et d'une baisse incroyable du rapport prix/rendement des ordinateurs. Les bureaux de statistiques ont grandement bénéficié de ce progrès en rationalisant leurs activités d'enquête. Nous pouvons maintenant prévoir une utilisation plus systématique de la nouvelle technologie dans des tâches nécessitant davantage "d'intelligence" que le calcul de base et le traitement des données inhérents à la production de statistiques. L'élaboration de statistiques de même que la conception de systèmes informatiques sont des exemples de ces tâches. La nouvelle technologie entraîne également des changements qualitatifs et structuraux autant dans les bureaux de statistiques que dans les rapports entre les bureaux de statistiques et leurs clients. Parmi les effets qui se manifestent déjà, notons (i) la standardisation de la technologie, des logiciels et de la méthodologie; (ii) la décentralisation des ordinateurs et des ressources para-informatiques; et (iii) l'intégration de différents types de tâches, de techniques et de compétences. Plusieurs bureaux de statistiques examinent d'un oeil critique la création de logiciels maison et on se demande si on verra de nouvelles générations de produits logiciels généraux spécialement destinés à la production de statistiques et à quoi ils ressembleront. La coopération internationale au chapitre de l'élaboration de logiciels entre les bureaux de statistiques pourrait constituer une solution intéressante. A l'extérieur des bureaux de statistiques, la nouvelle technologie offrira de nouvelles possibilités aux utilisateurs de statistiques et ceux-ci exigeront un meilleur service des producteurs de statistiques. Voilà qui nécessitera un réexamen et une réorganisation des bureaux de statistiques.

(1) ÉVOLUTION HISTORIQUE ET SITUATION ACTUELLE

Chacun sait que la technologie progresse à pas de géant et que les ordinateurs d'aujourd'hui sont très puissants et peu coûteux, aussi nous n'insisterons pas là-dessus. Il suffira de souligner que chacun des ordinateurs personnels que l'on retrouve aujourd'hui sur les bureaux du personnel d'un bureau de statistiques est à peu près aussi puissant que l'unité centrale du bureau de statistiques d'il y a 20 ans, et coûte une fraction du prix. Je crois qu'il serait également juste de dire que la plupart de ces ordinateurs personnels ne sont pas là pour symboliser un statut quelconque, mais qu'ils sont utilisés de façon très productive pour exécuter le travail du bureau de statistiques.

Il serait intéressant de comparer la situation actuelle avec certaines prévisions sur les besoins en informatique faites il y a quelques dizaines d'années lorsque cette technologie a fait son apparition et que, dans la pratique, elle est devenue disponible. A cette époque, certains des principaux informaticiens suédois croyaient vraiment qu'un seul ordinateur suffirait amplement aux besoins de notre pays dans un avenir prévisible. Des opinions similaires ont été émises dans d'autres pays. Ce type de pronostic est d'autant plus marquant car si l'on tient compte de la fonctionnalité (plutôt que de la puissance et du prix) des ordinateurs, le progrès n'a pas été si spectaculaire. Au niveau technique, les ordinateurs font essentiellement les mêmes choses aujourd'hui qu'à l'époque, sauf qu'ils le font avec beaucoup plus d'efficacité. Donc, en principe, il y a 30 ou 40 ans, les experts en informatique auraient dû être en mesure de prédire avec un peu plus de précision le potentiel de la technologie informatique. Mais tel n'a pas été le cas. Ils ont fait l'erreur de ne considérer qu'une seule catégorie d'application, soit les calculs mathématiques, et même là, ils n'ont pas prévu la myriade de besoins qui se manifesteraient une fois que la technologie serait disponible à grande échelle et à un prix minime.

Je tiens à dire que depuis les tous débuts de l'histoire de l'informatique, nous semblons avoir mis du temps à pouvoir apprécier pleinement les possibilités d'application de la technologie informatique, et à planifier activement l'utilisation la plus constructive de ces possibilités. Cela semble s'appliquer également aux bureaux de statistiques. Nous sommes pressés d'acquérir la technologie informatique de pointe mais, à mon avis, trop souvent nous employons beaucoup de nos ressources à transférer les mêmes vieilles applications d'une génération de technologie à l'autre plutôt qu'à développer activement de nouvelles applications, de nouvelles méthodologies et de nouvelles façons d'exécuter le travail global d'un bureau de statistiques, en nous appuyant sur des jugements stratégiques à long terme un peu plus imaginatifs sur la disponibilité future de la technologie informatique. De plus, nous semblons négliger les changements radicaux constants dans les rapports de coûts entre le matériel, les logiciels et les ressources humaines; nous oublions également de nous demander de façon formelle si des combinaisons plus ou moins nouvelles de ces facteurs de production ne seraient pas plus optimales. Par exemple, à Statistics Sweden, jusqu'à très récemment, l'acquisition d'un ordinateur personnel devait être formellement approuvée par le directeur général de notre agence tandis que l'on pouvait décider à un niveau de gestion inférieur d'engager une secrétaire, soit un investissement à peu près 1 000 fois plus important que l'achat d'un OP avec une période d'amortissement 10 fois plus longue.

(2) DIFFÉRENTS TYPES ET NIVEAUX DE SOUTIEN INFORMATIQUE

Dans une agence de statistiques, les possibilités d'utilisation d'ordinateurs sont très vastes. Certaines de ces possibilités ont déjà été exploitées dans une large mesure tandis que d'autres sont, au mieux, au stade de l'exploration. Afin de discuter du potentiel de la technologie moderne en ce qui a trait au travail statistique, nous avons besoin d'une restructuration fondamentale de ce travail. J'ai choisi d'utiliser deux structures différentes. La première consiste à classer les tâches statistiques en deux catégories : les tâches plus ou moins courantes et celles qui semblent nécessiter un peu plus "d'intelligence". L'autre classification nous amène à établir une distinction entre les opérations statistiques comme telles d'une part et le contrôle des opérations statistiques d'autre part, le "contrôle" englobant la planification, l'administration et l'évaluation. Parmi les tâches de contrôle, nous pouvons distinguer celles qui sont destinées aux enquêtes statistiques individuelles et celles qui s'appliquent à un système statistique entier tel le système statistique global d'un pays.

(2.1) APPLICATIONS COURANTES C. "INTELLIGENCE"

De nos jours, la grande majorité des applications sur ordinateur dans un bureau de statistiques sont de nature courante. Les données sont introduites, mises en forme, extraites, comptées et présentées plutôt simplement. Les calculs ne sont pas toujours très complexes, mais les volumes de données sont parfois assez importants. L'ordinateur n'est guère plus qu'un comptable extrêmement efficace qui ne fait aucune erreur. Néanmoins, cela a permis aux agences de statistiques d'économiser d'importantes sommes d'argent.

Cependant, le défi que nous devons maintenant relever est le suivant : pouvons-nous commencer à utiliser les ordinateurs de façon plus "intelligente"? Jusqu'à maintenant, nous avons très bien réussi à accroître la capacité de l'être humain à manipuler et à trier les données de même qu'à les compter, et à éliminer les risques d'erreurs humaines dans ces opérations. Mais pouvons-nous également utiliser l'ordinateur pour développer l'intelligence humaine sur le plan du travail statistique? Sans surestimer les possibilités de disciplines aux noms aussi évocateurs que "intelligence artificielle" et "systèmes experts", je crois que la production de statistiques offre de nombreuses occasions favorables où l'on pourrait utiliser des méthodes de systèmes experts. Nous en reparlerons plus en détail au chapitre 4.

(2.2) OPÉRATION D'ENQUÊTE C. PLANIFICATION D'ENQUÊTE, ADMINISTRATION ET ÉVALUATION C. PLANIFICATION STRATÉGIQUE DES TÂCHES ET ORGANISATION D'UNE AGENCE DE STATISTIQUES

Comme nous l'avons mentionné ci-dessus, il y a trois différents paliers de travail dans une agence de statistiques. Au premier palier, soit à la base, nous retrouvons les opérations statistiques concrètes qui constituent les parties opérationnelles d'une enquête statistique. Nous reconnaissons tous le déroulement traditionnel et séquentiel des tâches qui doivent être exécutées lorsque nous effectuons une enquête: collecte des données, codage, mise en forme, transformation des données, rassemblement, tabulation, présentation graphique, analyses, édition, distribution. Au chapitre 3, nous aborderons les possibilités existantes non utilisées de développement de soutien informatique.

Au deuxième palier, nous contrôlons les différentes étapes de l'enquête ainsi que l'enquête au complet au moyen de l'élaboration et de la planification, de l'administration et de l'évaluation. Le plan statistique comprend l'établissement d'une structure et d'une stratégie d'échantillonnage, s'il y a lieu, et la structure informatique comprend l'analyse des systèmes, la modélisation des données et la programmation. Le chapitre 4 traitera des possibilités d'amélioration de l'assistance informatique pour ces tâches.

Enfin, au troisième palier, c'est le système statistique dans l'ensemble qui fait l'objet du contrôle. Il peut s'agir du système d'information statistique d'un pays ou bien d'une partie d'un tel système géré par une agence de statistiques déterminée. Bien qu'un tel système repose sur de nombreuses enquêtes statistiques, il comprend également d'autres éléments et, dans le cas du système statistique comme dans celui des autres systèmes, l'ensemble ne doit pas se résumer à la somme de ses éléments. Au chapitre 6, nous verrons comment un système d'information statistique peut se développer au-delà de la somme des enquêtes qu'il comprend et comment une étude globale reposant sur la technologie moderne peut aider le super système à produire les résultats escomptés le plus efficacement possible.

(3) MAINTIEN DE LA RATIONALISATION DES OPÉRATIONS D'ENQUÊTE AU MOYEN DE LA STANDARDISATION, DE LA DÉCENTRALISATION ET DE L'INTÉGRATION

L'un des principaux messages que le présent document cherche à communiquer c'est que nous devrions chercher activement des possibilités d'utiliser la technologie moderne dans de nouveaux domaines et pour de nouveaux aspects de la production statistique au lieu de nous limiter à transférer "les mêmes vieilles applications" d'une génération d'ordinateur à une autre. Par ailleurs, nous ne devons pas négliger les possibilités de faire "les mêmes choses" de façon plus efficace en appliquant la nouvelle technologie et la nouvelle méthodologie aux tâches traditionnelles que suppose l'exécution d'une enquête statistique. J'utiliserai trois mots pour décrire ce qui doit être fait : standardisation, décentralisation et intégration.

(3.1) STANDARDISATION DE LA TECHNOLOGIE, DES LOGICIELS ET DE LA MÉTHODOLOGIE

Les ordinateurs sont maintenant si bon marché et le personnel si coûteux qu'il vaut rarement la peine de rechercher l'efficacité technique maximale dans la conception d'un système informatisé. Bien entendu, dans le cas d'un gros système avec des volumes importants de données à traiter et supposant un nombre considérable d'interactions homme-machine, il peut parfois être nécessaire, ou du moins plus rentable, de mobiliser l'ingéniosité technique afin d'éliminer les éventuels goulets d'étranglement et d'accélérer le temps de réponse, ou d'économiser de la capacité de mémoire et de traitement. Mais nous devons d'abord faire remarquer deux choses. Premièrement, il n'existe pas un si grand nombre de systèmes de ce genre, même pas dans un bureau de statistiques comprenant d'importantes bases de données. La plupart des enquêtes statistiques sont plus ou moins importantes, elles sont traitées plutôt rarement et les exigences relatives au temps de réponse sont souvent très modérées comparativement à celles d'un grand nombre de systèmes commerciaux interactifs de type administratif.

L'autre remarque que je tenais à faire à cet égard, c'est que même dans les rares cas où du point de vue administratif, l'optimisation technique est vraiment optimale, l'optimum administratif ne consiste habituellement pas à optimiser techniquement l'ensemble du système mais seulement quelques parties ou aspects limités de celui-ci.

Devant ces observations, lorsqu'il s'agit de la rationalité de l'optimisation technique, le fardeau de la preuve devrait toujours revenir à celui qui prétend que c'est nécessaire. La solution par défaut devrait toujours être simple, directe et normalisée.

Ce que je viens de dire peut paraître si évident qu'il était superflu de le mentionner. Malheureusement, cette "vérité de La Palice" ne semble pas toujours bien comprise ni acceptée dans les bureaux de statistiques. Même si, en principe, nous acceptons sa validité, il semble que nous négligions ses conséquences dans la pratique, entre autres dans la pratique administrative. Cela semble être une occasion pour tous les intervenants-programmeurs, analystes fonctionnels, et gestionnaires, d'employer tous leurs efforts à perdre une vieille habitude puri-

taine (ne gaspiller aucun type de ressources matérielles sous aucune considération) au profit de l'économie et de l'utilité globale des systèmes conçus.

Autrement dit, aucun analyste fonctionnel ou programmeur ne devrait s'écarter des normes établies par les cadres supérieurs d'un bureau de statistiques sans en avoir préalablement parlé à un gestionnaire responsable. De même, aucun gestionnaire responsable ne devrait accepter d'écart à la norme sans arguments valables étayés par une bonne documentation dont il/elle reconnaît pleinement le contenu. Les vérificateurs internes et externes devraient considérer toute violation à ces règles comme une faute grave.

La règle de gestion proposée admet qu'au sein de l'agence de statistiques, une norme bien établie, ou politique, régit des aspects importants et tangibles de la conception de systèmes informatiques. Idéalement, cette politique devrait être intégrée à une méthodologie théoriquement sûre en matière de conception et de mise en application de systèmes, de façon que quiconque utilise cette méthodologie et ses outils de travail (voir section 4.2) suive plus ou moins automatiquement les règles de la politique. Le matériel, les logiciels, les interfaces et la méthodologie elle-même, y compris les règles sur la documentation, devraient faire l'objet d'une standardisation.

Standardisation du matériel et des systèmes d'exploitation

Pour le moment, la standardisation du matériel et des systèmes d'exploitation ne constitue pas réellement un problème majeur puisque l'industrie l'a résolue (par chance?) pour les utilisateurs en établissant des normes de fait très strictes: des processeurs centraux compatibles IBM, des micro-ordinateurs compatibles IBM, possiblement des systèmes d'exploitation UNIX. Cependant, il ne faudrait pas oublier que cette situation favorable n'est pas du tout attribuable à une intervention précise des utilisateurs. Par conséquent, il pourrait y avoir un retour assez rapide à la traditionnelle absence de normes.

Standardisation des logiciels

Dans le domaine du logiciel, il importe de souligner que l'on semble avoir enfin accepté le fait que les utilisateurs ne devraient pas élaborer de logiciels propres à leurs applications. Ils devraient plutôt utiliser les logiciels généraux et, s'il y a lieu, adapter ces produits à leurs applications particulières. En conséquence, le nombre de programmeurs d'applications devrait diminuer considérablement dans les bureaux de statistiques, si ce n'est pas déjà le cas. Certains bureaux de statistiques ont même commencé à mettre en question leur propre élaboration de logiciels généraux. Donc, il semble que nous ne soyons pas loin de mettre en question toute la catégorie professionnelle de programmeurs en tant que groupe de spécialistes identifiables dans les bureaux de statistiques. Nous reviendrons sur cette question controversée.

Il serait bon de noter que, dans un bureau de statistiques, lorsque nous parlons de logiciel standardisé ou général, nous faisons référence à deux importantes sous-classes. Le produit qui a été mis au point pour des tâches statistiques particulières, ou des fonctions caractéristiques d'un bureau de statistiques, constitue un type de logiciel général. Un tel logiciel statistique général, ou logiciel général spécialisé, est souvent élaboré par les bureaux de statistiques eux-mêmes ou par des instituts de recherche statistique. Toutefois, il existe un autre type de logiciel général que nous pouvons appeler logiciel général à usages multiples. Un tel produit est

ordinairement élaboré à des fins commerciales générales et non pour des applications statistiques particulières, quoique qu'il puisse également s'avérer utile dans des contextes statistiques. Les systèmes de gestion de base de données en sont un bon exemple.

Méthodologie standard

Aujourd'hui, en Suède, la plupart des grandes compagnies et organisations informatisées ont adopté des méthodologies standard pour le développement et la maintenance des systèmes informatiques. Les méthodologies portent le nom de modèles de développement de systèmes, et indiquent quelles tâches devraient être exécutées pendant le développement des systèmes, dans quel ordre elles devraient l'être, quels concepts utiliser pendant l'analyse, comment visualiser les concepts et les résultats de l'analyse, quelles règles et normes doivent être observées, quels sont les documents à préparer, etc. Habituellement, on retrouve une norme ou du moins une variation d'une norme par compagnie. Cependant, une sorte de norme de fait semble vouloir ressortir, même d'une compagnie à l'autre, comprenant des caractéristiques telles que la distinction entre les phases "infologiques" (axées sur le contenu) et "datalogiques" (axées sur la technique), l'accent étant mis sur la modélisation conceptuelle selon une "méthodologie des trois concepts" (objets, biens, rapports ou entités, attributs, rapports) durant la phase infologique, et sur le modèle de données relationnelles pendant la phase datalogique. Pour compléter la modélisation des concepts et des données orientés vers l'état, les méthodologies de développement de systèmes contiennent souvent des techniques de modélisation axées sur le déroulement et d'autres méthodes permettant de clarifier les aspects dynamiques du système.

L'un des aspects importants des méthodologies de développement des systèmes c'est qu'elles peuvent prévoir des interfaces normalisées, par exemple des formats standard pour la mémorisation et la transmission de données et de métadonnées entre les différentes parties d'un système et des interfaces standardisées pour les utilisateurs en général, ou des syntaxes standard pour les langages des utilisateurs, en particulier.

La méthodologie de développement de systèmes de Statistics Sweden se nomme modèle SCB et est utilisée depuis plus de dix ans. Elle a également été une source d'inspiration pour les modèles de développement de systèmes de plusieurs compagnies et organisations (non statistique) importantes en dehors de Statistics Sweden.

(3.2) DÉCENTRALISATION DES RESSOURCES INFORMATIQUES ET PARA-INFORMATIQUES

L'élimination du besoin de contrôle centralisé des ressources informatiques compte parmi les conséquences organisationnelles évidentes des changements au niveau des rapports de coûts. Nous pouvons maintenant acheter ces ressources en petites quantités et à très bas prix. Ainsi, il n'est plus nécessaire de centraliser pour partager des ressources coûteuses, indivisibles et rares. A la place, nous pouvons intégrer les décisions concernant les ressources informatiques à d'autres décisions importantes du bureau de statistiques et tenter d'élaborer le même genre de "décentralisation équilibrée" des prises de décisions que l'on retrouve dans d'autres domaines, en laissant les personnes chargées des enquêtes statistiques assumer l'entière responsabilité de tous les types de ressources nécessaires à l'élaboration et au déroulement de l'enquête, sans traiter les ressources informatiques d'une façon particulière.

En Suède, nous avons adopté d'importantes mesures en ce sens. Le gouvernement a engagé un processus selon lequel les agences possédant suffisamment de compétences et d'expertise en traitement électronique des données peuvent prendre la plupart des décisions reliées à l'informatique sans avoir à consulter une autre agence ou le ministère des Finances. Bien sûr, les agences jouissent de cette liberté dans la mesure où elles peuvent assumer toutes leurs décisions dans les limites d'un budget donné.

A Statistics Sweden, nous tentons également d'intégrer les décisions reliées au traitement électronique des données à d'autres décisions. Nous avons établi une politique d'informatique qui évidemment sera mise à jour de temps à autre et qui permettra à chaque service de prendre ses propres décisions dans la mesure où il respecte son budget. Ainsi, tous les types de coûts - matériel, logiciels, personnel, ressources apparentées au processeur central et aux micro-ordinateurs - sont, dans la mesure du possible, évalués en fonction des "mêmes types de fonds".

Dans le domaine de la conception et de la programmation de systèmes d'application, nous avons amorcé le processus de décentralisation il y a déjà cinq ans. La responsabilité de même que les ressources humaines pour ces activités sont transférées du service central des systèmes aux différentes divisions. Les sections du service des systèmes qui restaient et d'autres fonctions de développement ont fusionné pour former une nouvelle division de recherche et développement et ont conservé des fonctions telles que l'élaboration et la maintenance de logiciels généraux, la recherche, et le développement dans le domaine du traitement des données statistiques et la formation en informatique.

Le processus de décentralisation s'est poursuivi plus ou moins rapidement à l'intérieur de chaque division. Dans certains cas, il existe encore des groupes relativement importants d'analystes fonctionnels et de programmeurs, et dans d'autres, le processus de décentralisation s'est poursuivi jusqu'au niveau des unités et des enquêtes consacrées à des sujets individuels.

Cette décentralisation s'est avérée positive en ce sens que le gestionnaire d'une enquête statistique déterminée possède maintenant une vue d'ensemble, une connaissance et un contrôle beaucoup plus complets de son produit et de tous les types de ressources requises, en supposant naturellement qu'il a les capacités et la bonne volonté nécessaires pour utiliser ces possibilités. Par ailleurs, il y a toujours le risque de voir le bureau de statistiques se subdiviser en de nombreuses petites organisations non coordonnées basées sur les enquêtes. Afin d'éviter cela, on a créé un certain nombre de "conseils" spécialisés (dont un pour l'informatique) chargés de donner à la haute direction et au directeur général du bureau de l'information spécialisée en matière de politiques, etc.

A mon avis, ce processus de décentralisation a été très réussi en général; il était nécessaire et est loin d'être terminé. Espérons qu'il mènera, entre autres choses, à une meilleure intégration des différents aspects méthodologiques de la production de statistiques, y compris l'intégration de la méthodologie statistique et le traitement électronique des données. Nous reviendrons à cette question dans la prochaine section.

(3.3) INTÉGRATION DE DIFFÉRENTS TYPES DE TÂCHES, TECHNIQUES ET COMPÉTENCES

L'intégration de différentes tâches, techniques et compétences constitue l'envers de la "médaille de la décentralisation". Les effets de cette intégration commencent à se manifester à travers l'organisation. Les gestionnaires perdent leurs secrétaires particulières et leurs adjoints administratifs, remplacés par des ordinateurs personnels, des réseaux, et des systèmes de bureautique. Les statisticiens chargés de divers sujets s'occupent du développement et de la maintenance des applications sans l'aide des analystes fonctionnels et des programmeurs et font leur propre édition sans avoir recours aux typographes. Les enquêteurs assument les tâches d'introduction et de mise en forme des données.

La décentralisation des ressources informatiques et para-informatiques a permis et facilité l'intégration de ces tâches et d'autres tâches similaires. L'intégration a plusieurs effets positifs dont l'extension des emplois, l'abrégement des cheminements de communication et de décision, la réduction des frais généraux et la répartition plus nette des responsabilités. Cependant, "l'amateurisme insouciant" risque de remplacer le professionnalisme compétent et efficace, et l'isolationnisme et la suffisance peuvent s'installer au sein des unités organisationnelles relativement petites et indépendantes. Mais, dans l'ensemble, les effets positifs semblent l'emporter sur les effets négatifs et les membres du personnel ayant l'impression d'avoir tiré profit de ce développement sont plus nombreux que ceux qui croient avoir perdu. Même quelques-uns de ceux qui ont perdu des responsabilités et des pouvoirs reconnaissent ou acceptent le développement comme étant fondamentalement sain et se trouvent de nouveaux rôles relativement rapidement au sein de l'organisation.

Bien sûr, les bureaux de statistiques auront toujours besoin de bons spécialistes dans plusieurs domaines. Le progrès technologique continu ne fait qu'éliminer le besoin de centralisation et de "fonctionnalisation", lequel reposait sur l'indivisibilité des gros ordinateurs coûteux et la rareté relative des analystes fonctionnels et des programmeurs. Toutefois, le développement a également soulevé la question selon laquelle il ne serait pas rationnel de viser un degré plus élevé d'intégration entre les disciplines pertinentes pour la production de statistiques. L'actuelle répartition en spécialités que l'on retrouve dans les bureaux de statistiques reflète généralement l'organisation universitaire. Pour un bureau de statistiques, une meilleure collaboration entre les experts en méthodologie statistique et informaticiens serait nettement plus avantageuse. Une telle intégration serait encore plus importante si elle pouvait également amener un changement au niveau des universités, faisant de la production de statistiques en soi un domaine établi et respecté de recherche et de formation universitaires.

(4) CONCEPTION ASSISTÉE PAR ORDINATEUR D'ENQUÊTES STATISTIQUES ET SYSTÈMES D'INFORMATION STATISTIQUE

Nul doute que la prochaine vague d'informatisation des agences de statistiques touchera le milieu des opérations statistiques plutôt que les opérations statistiques comme telles. Dans l'ensemble, nous avons déjà informatisé la manipulation des données de base ainsi que les calculs inhérents à l'exécution d'une enquête statistique. Il est maintenant temps d'envisager sérieusement une informatisation massive des tâches reliées au contrôle d'une enquête statistique, notamment la planification et l'élaboration de l'enquête, d'un point de vue statistique, la conception et la construction du système de production (qui est en grande partie un système informatisé), la gestion des activités de production et l'évaluation de la performance de l'enquête statistique. Dans ce chapitre, nous étudierons certaines de ces tâches ainsi que leur potentiel en termes d'informatisation.

(4.1) PLAN STATISTIQUE

Dans le milieu statistique, on sait bien que le plan statistique est essentiel à l'organisation et à l'exécution d'une enquête statistique. Je suis également certain que la plupart des experts en méthodologie statistique ont déjà reconnu l'efficacité de l'ordinateur comme machine à calculer et outil de travail. On peut se demander si l'informatisation du plan statistique peut se faire d'une façon plus systématique et tendre davantage vers l'atteinte d'un but. Seuls les statisticiens, de concert avec les informaticiens, peuvent répondre à cette question.

Personnellement, je crois que la méthodologie statistique appuyée par une utilisation intelligente des ordinateurs pourrait produire une autre vague de rationalisation de la production de statistiques possiblement de la même ampleur que ce que les ordinateurs ont déjà accompli, seuls. Il semble que les non-statisticiens considèrent parfois la méthodologie statistique comme un "mal nécessaire" que l'on doit supporter pour assurer une qualité raisonnable des résultats d'enquête statistique et pour éviter le mauvais usage, conscient ou non, des statistiques. Évidemment, il s'agit d'une importante fonction de la méthodologie statistique mais celle-ci, combinée à la technologie moderne, pourrait également s'avérer une force active extrêmement puissante aux niveaux de la rationalisation et de l'économie dans la production de statistiques.

Permettez-moi de citer quelques exemples. Dans une importante enquête sur les ménages suédois, une équipe d'experts en méthodologie statistique, aidée d'analyses informatisées et de simulations, propose une stratégie plus efficace d'échantillonnage et de stratification. Les résultats de cet exercice pourrait servir à améliorer la qualité des estimations produites dans l'enquête. Ou bien, la taille de l'échantillon pourrait être réduite d'environ 50%, diminuant de moitié le coût de collecte des données, un article de budget important (environ 2 millions de couronnes) dans cette enquête.

Les gains considérables qui peuvent souvent être réalisés grâce à une utilisation combinée bien élaborée d'enquêtes par sondage et de registres administratifs constituent un autre

exemple. Dans les pays scandinaves, ce plan pourrait éliminer le besoin de recensement, permettant ainsi d'économiser d'importantes sommes d'argent. Cette méthode offre une particularité intéressante : elle semble améliorer la qualité du registre et de l'enquête tout en réduisant les coûts.

Bien entendu, les experts en méthodologie statistique pourraient vous fournir une longue liste d'exemples encore plus précis. Toutefois, ces derniers ne semblent pas afficher ni vendre très souvent de cette façon le pouvoir de rationalisation de leurs outils méthodologiques. En conséquence, il me semble que trop de statisticiens se tournent trop souvent vers les toutes dernières nouveautés en technologie informatique au lieu d'utiliser sérieusement certaines des possibilités de rationalisation que leur offre la méthodologie statistique appuyée par ordinateur.

(4.2) CONCEPTION DE SYSTÈME

L'usage croissant de logiciels généraux au lieu des programmes d'application personnalisés a grandement amélioré l'efficacité au niveau de la construction et de la maintenance des systèmes. Toutefois, l'usage de logiciels généraux ne réduit pas le besoin d'une bonne conception de système. Cela ne diminue pas beaucoup non plus la somme de travail des spécialistes compétents pendant les premières phases de la conception. Il peut sembler surprenant que ces mêmes spécialistes, qui sont responsables de la conception d'un si grand nombre de systèmes informatisés n'utilisent pas davantage les systèmes informatisés pour appuyer leurs efforts. Après tout, nous sommes depuis longtemps habitués à des concepts tels que Enseignement assisté par ordinateur (EAO), Conception assistée par ordinateur (CAO) et Fabrication assistée par ordinateur (FAO).

Cependant, il se passe quelque chose de nouveau dans ce domaine. Un nouvel acronyme vient de voir le jour : il s'agit de CASE qui signifie ingénierie logicielle assistée par ordinateur. La boîte à outils du CASE renferme des instruments logiciels pour soutenir les différentes étapes de travail dans les modèles de développement de systèmes. Le problème c'est que chaque organisation possède son propre modèle de développement de systèmes et qu'il n'existait pas nécessairement sur le marché de boîtes à outils CASE parfaitement adaptées aux besoins de ce modèle. L'organisation se trouve donc devant une alternative : devoir changer sa méthodologie de développement de systèmes ou élaborer sa propre boîte à outils CASE.

Comme je l'ai mentionné à la section 3.1, dans le domaine des modèles de développement de systèmes, il existe un procédé de standardisation de fait. Pour une organisation, ce procédé, combiné aux possibilités d'acquérir et d'adapter un système expert générique CASE plutôt que d'accepter tous les détails d'un outil CASE déjà prêt, devrait aider à résoudre le problème.

Ce développement devrait présenter énormément d'intérêt pour les bureaux de statistiques. Tout en attendant qu'un système expert générique CASE adéquat soit mis sur le marché, l'agence de statistiques pourrait elle-même prendre un certain nombre de mesures relativement simples pour améliorer le soutien informatique au niveau de la conception de système.

L'élaboration d'un outil interactif pour la création et la mise à jour de documentation sur les systèmes apparaît évidente mais elle constitue néanmoins une mesure importante.

De plus, il importe d'avoir un outil qui transforme et communique automatiquement les métadonnées entre les différents produits logiciels.

En outre, on pourrait se demander si le concepteur/l'utilisateur devrait s'occuper des différences techniques plus ou moins importantes entre les divers progiciels ou même de la sélection en premier lieu d'un produit logiciel particulier (plutôt qu'un autre). Idéalement, le concepteur/l'utilisateur devrait uniquement préciser la fonction (par exemple, tabulation) qu'il ou elle veut voir exécuter et l'outil de développement de systèmes devrait automatiquement choisir (ou proposer) un produit logiciel et générer une application complète d'après les métadonnées provenant du système de documentation et les quelques données entrées par le concepteur/l'utilisateur pour exprimer ses préférences sur certaines questions.

A Statistics Sweden, on a mis au point un outil de développement de systèmes et de documentation se rapprochant des idéaux mentionnés ci-dessus. On l'appelle CONDUCTOR et il execute présentement sur le processeur central. Il accélère même le travail des analystes fonctionnels expérimentés et permet à ceux qui ne sont pas des professionnels de l'informatique de mettre leurs propres applications en service, pourvu que leur problème soit relativement simple et(ou) qu'ils comprennent suffisamment les premières phases du modèle de développement des systèmes axées sur le contenu.

(4.3) MÉTHODES AXÉES SUR LES CONNAISSANCES

De nos jours, on utilise souvent l'expression "système expert" à la place de "intelligence artificielle". Le profane serait en droit de se demander pourquoi nous devons subitement parler de "systèmes experts" et de méthodes axées sur les "connaissances" en production de statistiques. (N'avons-nous pas toujours eu recours à des méthodes axées sur les connaissances? Quelles autres méthodes pourraient exister?) Il est vrai, nous l'espérons, que la production de statistiques a toujours été axée sur la connaissance mais lorsqu'il nous venait à l'idée de consigner ces connaissances, nous le faisions habituellement sur papier, séparément des fichiers informatisés contenant les données traitées conformément à ces connaissances. Et enfin, si les connaissances étaient dans une certaine mesure représentées dans un ordinateur, elles étaient habituellement stockées de façon implicite dans les programmes.

Par contraste, la méthodologie informatique axée sur les connaissances suppose que:

- les connaissances utilisées dans les différentes étapes de la production statistique sont informatisées (du moins partiellement);
- les connaissances sont regroupées par faits et règles dans une base de connaissances, traitée conformément aux principes des bases de données;
- l'exploitation des connaissances est activement appuyée par ordinateur.

Ainsi, bien que le principe soit discutable, on doit admettre que l'application des méthodes axées sur les connaissances pour la production de statistiques vise à capter au moins des petites bribes de connaissances qui, jusqu'à maintenant, étaient considérées comme étant réservées aux statisticiens, et à mettre cette connaissance à la disposition des ordinateurs et des utilisateurs d'ordinateurs.

Nous serions en droit de mettre en doute le réalisme de cet objectif, mais je crois que nous ne pouvons nous permettre de rejeter catégoriquement ces idées et propositions. Nous devons comprendre que cette "explosion" de données dans la société ne fait que commencer. N'importe qui sera bientôt en mesure techniquement de produire des statistiques à partir de ces données. Comment pouvons-nous empêcher les interprétations erronées et le mauvais emploi de cette production de statistiques? Naturellement, le mieux serait que chaque producteur amateur de statistiques demande conseil à un statisticien qualifié et expérimenté. Mais même s'il y avait assez de statisticiens pour pouvoir le faire, je ne suis pas certain que la plupart des utilisateurs suivraient cette voie et je ne suis pas certain non plus que les statisticiens qualifiés et expérimentés aimeraient consacrer 99% de leur temps à donner de menus conseils sur des problèmes statistiques courants.

Je pense que si nous raisonnons dans ce sens, nous pouvons facilement nous accorder sur un rôle fondé, souhaitable et réaliste pour la méthodologie informatique axée sur les connaissances en matière de production de statistiques. Dans cette perspective, les systèmes experts constituent une phase normale du développement, après le système de métadonnées et les interfaces interactives pour utilisateurs (voir le système CONDUCTOR mentionné ci-dessus) qui sont déjà productifs et que la plupart d'entre nous apprécions.

En outre, je suis plutôt convaincu que les efforts en vue d'élaborer des systèmes informatisés axés sur les connaissances pour la production de statistiques auront des répercussions très positives, même si les objectifs les plus ambitieux ne seront pas faciles à atteindre. Je crois que nous sommes tous d'avis que le personnel, avec ses compétences et ses connaissances, constitue l'actif le plus important d'une agence de statistiques. Lorsque les spécialistes prennent leur retraite ou lors de compressions budgétaires, nous avons beaucoup de mal à conserver ce capital de connaissances. La documentation systématique des connaissances (appelée acquisition des connaissances dans le jargon des systèmes experts) devrait atténuer ces problèmes et fournir d'excellents instruments pour la formation interne.

(5) LA PROCHAINE GÉNÉRATION DE LOGICIELS STATISTIQUES

Chaque fois que l'on entame une discussion sur la prochaine génération de logiciels statistiques, on se demande en premier lieu s'il y aura effectivement une prochaine génération de logiciels statistiques, du moins si nous entrevoyons des produits logiciels élaborés par les bureaux de statistiques eux-mêmes. Certains bureaux de statistiques ont déjà commencé à mettre en question le besoin de programmeurs et de création de logiciels internes.

Je ne suis certainement pas le seul qui verrait mal un bureau de statistiques sans une certaine compétence en matière de création de logiciels évolués. Un désengagement dans ce domaine constitue probablement un processus irréversible qui aurait des conséquences négatives. Par exemple, il peut s'avérer difficile d'évaluer, de choisir, d'installer et d'adapter un logiciel commercial pour répondre aux besoins spécifiques d'un bureau de statistiques si ce bureau n'est pas doté de personnel compétent en matière de création de logiciels.

Par ailleurs, on doit admettre qu'il sera de plus en plus difficile pour les bureaux de statistiques de justifier de prestigieux projets de création de logiciels très coûteux comme ceux que nous avions l'habitude de lancer dans le passé et que nous réussissions parfois (mais pas toujours) à compléter et à mettre en application.

Nous pouvons prendre certaines mesures afin de résoudre cette situation difficile. L'une d'entre elles consiste à établir une politique et une structure de base pour la création de logiciels, en garantissant des caractéristiques importantes telles que les possibilités de cohabitation entre les progiciels commerciaux et les éléments de logiciels maison, la modularité et l'incrémentation dans la création de logiciels et l'interchangeabilité entre les différents types d'ordinateurs.

Une autre mesure possible, qui devrait être combinée à la première, consiste à compter davantage sur la coopération internationale entre les agences de statistiques. J'utiliserai quelques-unes des expériences du Projet de calcul statistique (PCS) de ONU/Commission économique pour l'Europe comme base de discussion sur ces sujets.

(5.1) COOPÉRATION INTERNATIONALE: LE PROJET DE CALCUL STATISTIQUE (PCS) DE LES NATIONS UNIES ET LA COMMISSION ÉCONOMIQUE POUR L'EUROPE (ONU)

Le sigle PCS symbolise un effort de coopération dans le domaine des calculs statistiques entre les pays membres de la Commission économique pour l'Europe (ONU), comprenant les pays d'Europe, le Canada et les États-Unis. Le PCS s'est déroulé sous diverses formes et avec des thèmes variés de coopération depuis les années 1980. Il a d'abord vu le jour comme projet (PCS-1) appuyé par l'UNDP dans une certaine mesure, puis, lors de la Conférence des statisticiens européens, il est devenu un programme (PCS) et depuis environ un an demi, il est redevenu un projet (PCS-2) appuyé par l'UNDP. Tandis que le PCS-1 était principalement axé sur

le processeur central, le PCS-2 devrait porter tout particulièrement sur l'usage croissant des micro-ordinateurs dans les bureaux de statistiques.

Six groupes mixtes se sont partagés le travail considérable relié au PCS-2.

- Groupe mixte chargé de l'évaluation des logiciels (pays responsable: Hongrie)
- Groupe mixte chargé de la communication (pays responsable: France)
- Groupe mixte chargé de la stratégie de mise en application (pays responsable: Pologne)
- Groupe mixte chargé de la gestion de bases de données statistiques (pays responsable: Suède)
- Groupe mixte chargé de la mise en forme des données (pays responsable: Yougoslavie)
- Groupe mixte chargé de la génération des tables (pays responsable: République démocratique allemande)

Les activités des groupes mixtes sont contrôlées par une équipe de travail composée des pays responsables des groupes mixtes et cette équipe relève d'un comité directeur constitué de tous les pays participant au PCS-2. Les trois derniers groupes mixtes de la liste ci-dessus sont engagés activement dans la création de logiciels dans leur domaine d'intérêt respectif.

(5.2) QUELQUES CARACTÉRISTIQUES SOUHAITABLES DES LOGICIELS

Pour tenter d'augmenter l'efficacité totale de l'élaboration de logiciels à laquelle travaillent actuellement les différents groupes mixtes, l'équipe de travail du PCS a demandé à l'auteur de ce document de présenter des propositions de coordination concrètes. Idéalement, cet effort devrait donner ce qui suit:

- une description d'une méthode de conception unifiée ainsi qu'une structure de logiciels que peuvent se partager tous les groupes mixtes dans le cadre de l'élaboration continue de logiciels
- une proposition en faveur d'une méthode uniforme de traitement des métadonnées dans les produits logiciels PCS
- une proposition en vue de faciliter l'échange de données entre les différents produits logiciels (logiciels PCS, progiciels commerciaux, produits "maison", etc.) et entre les différents ordinateurs
- une proposition en vue d'assurer la transférabilité du logiciel PCS entre les différents types d'ordinateurs (micro/ mini/processeur central)
- une suggestion de mesures à prendre pour faciliter la "commercialisation" des produits PCS comme membres d'une seule et même famille de logiciels

- un plan provisoire et synchronisé des activités devant être exécutées par chaque groupe mixte pour atteindre les objectifs communs de l'élaboration du logiciel PCS

Certaines des propositions issues de ce mandat seront présentées dans la prochaine section.

(5.3) STRUCTURE DE LOGICIEL PROPOSÉE

Le rapport présenté à la suite de l'effort coordonné sus-énoncé comprend les propositions suivantes:

1. Toute l'élaboration de logiciels PCS devrait reposer sur un modèle de production de statistiques axé sur la base de données. Il importe de bien définir les opérations élémentaires de ce modèle et les définitions devraient s'appuyer sur une analyse fonctionnelle logiquement rigoureuse des principales fonctions types en production de statistiques. Toutes les opérations intermédiaires de la chaîne de production devraient utiliser et produire le même type d'éléments de base de données : tables à deux dimensions ou relations au sens du modèle de données relationnelles. Par conséquent, au sens mathématique, avec ce type de structure de données les opérations constitueraient une algèbre. Cette base conceptuelle pour l'élaboration de logiciels PCS assurera modularité, simplicité, facilité de combinaison et capacité de développement incrémentiel.

2. Outre des éléments de logiciel de base transférables et bien intégrés pour la mise en forme, la gestion de base de données et la tabulation, le progiciel PCS devrait comprendre un cadre de développement de systèmes (CDS) facile à utiliser, composé d'une interface interactive d'utilisateur (IIU) et, si possible, de certains outils reposant sur les principes de l'ingénierie logicielle assistée par ordinateur et de la méthodologie axée sur les connaissances (systèmes experts). L'IIU devrait, entre autres choses, aider l'utilisateur à venir à bout de toute différence pouvant exister au niveau des langages entre les divers éléments du logiciel PCS et qui existeront certainement entre le progiciel PCS et d'autres produits logiciels que l'utilisateur pourrait vouloir combiner au logiciel PCS. Par conséquent, l'IIU devrait aider à standardiser l'interface de l'utilisateur malgré les inévitables différences entre les produits logiciels. Par ailleurs, l'IIU pourrait également servir à individualiser (personnaliser) l'interface de l'utilisateur pour répondre aux besoins spécifiques d'un groupe d'utilisateurs déterminé ou d'un bureau de statistiques en particulier.

3. L'uniformité des produits logiciels PCS au niveau conceptuel devrait trouver son équivalent naturel au niveau technique. On a déjà mis en application une algèbre des opérateurs travaillant aux structures de données standardisées (tables à deux dimensions et relations) et ces principes devraient s'étendre à d'autres projets de logiciels PCS. On propose ici un compilateur comme équivalent technique des opérateurs de l'algèbre. Chaque élément logiciel PCS devrait être conçu en fonction des compilateurs, et des compilateurs exécutant des tâches identiques ou similaires devraient être standardisés et mis en application une seule fois. De façon identique, on pourrait utiliser le même compilateur dans plusieurs parties du même élément de logiciel et dans plusieurs éléments de logiciel.

4. Le langage du compilateur, qui contrôle les opérations des compilateurs, pourrait parfois être le langage de l'utilisateur. C'est le cas, par exemple, du Système opérateur de base. Pour des fonctions plus complexes telles que la mise en forme et la tabulation, il est bon d'avoir un langage d'utilisateur spécial, traduit dans le langage du compilateur.

5. On propose la mise en application d'une méthode d'accès relationnel (MAR) commune en vue de standardiser et de faciliter l'accès et la transmission de données dans les compilateurs, entre ceux-ci, entre les produits logiciels PCS et entre les produits PCS et l'extérieur. Tous les compilateurs de tous les produits logiciels PCS devraient utiliser cette méthode d'accès pour la lecture et l'enregistrement des données. La MAR devrait consister en un ensemble de macros indépendantes pouvant être intégrées dans les modules du compilateur. Pour pouvoir relier un progiciel externe à la MAR, il faudra élaborer les macros de lecture/d'enregistrement adéquates, etc. pour le progiciel déterminé, une tâche relativement simple toutefois.

6. De la même façon, pour standardiser et faciliter le traitement des métadonnées, y compris - dans la mesure du possible - le transfert automatique et la transformation des métadonnées entre les compilateurs, entre les produits logiciels PCS et entre les logiciels PCS et les progiciels externes, on propose la mise en application d'un système commun de gestion des métadonnées. Le système de gestion des métadonnées devrait être utilisé par tous les compilateurs de tous les logiciels PCS et devrait être composé d'un ensemble de macros pour la lecture, l'enregistrement, la mise à jour, la suppression, etc. des métadonnées. L'utilisation d'un tel ensemble de macros faciliterait le soutien des différents types et formats de métadonnées sans qu'il soit nécessaire d'apporter des changements aux compilateurs. Toutefois, il convient de noter qu'il s'agit d'une proposition de standardisation du traitement des métadonnées et non de standardisation de tous les types de métadonnées, qui ne semble pas être un objectif réaliste à l'heure actuelle.

7. Afin d'assurer la transférabilité maximale des logiciels PCS entre les différentes catégories d'ordinateurs et de systèmes d'exploitation, on propose d'utiliser le langage de programmation C pour toute élaboration de logiciel. La transférabilité devrait (avec la priorité 1) être assurée pour les micro-ordinateurs compatibles avec le "PC" d'IBM fonctionnant avec le système d'exploitation à disques (SED) et le système d'exploitation pour micros 16 bits, de Microsoft inc., pour les processeurs centraux compatibles avec le IBM 370 et fonctionnant avec le système d'exploitation OS/MVS et VM/CMS, et pour le système d'exploitation UNIX.

8. La conception, la mise en application et la documentation de tous les produits logiciels PCS devraient englober les aspects suivants, dans l'ordre présenté:

   - analyse fonctionnelle et description explicite de la fonction de production statistique à l'étude

   - élaboration d'un manuel de référence pour le langage de l'utilisateur final

   - spécification sur le module en langage symbolique

   - élaboration d'un manuel sur le système logiciel à l'étude et comprenant une description

     — de la structure logique du programme

— des instructions sur le langage du compilateur utilisées et produites par le logiciel

— des possibilités d'enregistrement de sorties sur mesure

— des autres caractéristiques permettant une utilisation poussée du logiciel

- codage en langage C

- préparation:

  — d'un guide de l'utilisateur, articulé autour d'exemples pédagogiques

  — d'un guide d'installation

  — d'une bande et de disquettes d'installation comprenant le logiciel et sa documentation

  — de toute documentation supplémentaire, s'il y a lieu

9. Lorsque cela s'applique, la conception, la mise en application et la documentation des éléments communs des logiciels PCS, tels que le langage du compilateur, la méthode d'accès relationnel et le système de gestion des métadonnées devraient englober les mêmes articles que ceux qui sont mentionnés ci-dessus.

COMMENTAIRES SUR LES PROPOSITIONS

Les propositions présentées ci-dessus sont relativement ambitieuses et bien entendu, il reste à savoir dans quelle mesure elles seront acceptées et réalisées par les groupes mixtes du PCS. On remarquera plus particulièrement que les propositions supposent que l'élaboration des logiciels sera effectué dans le langage de programmation C. Comme autre solution, on pourrait baser la majeure partie de l'élaboration de logiciels sur des techniques de génération de programmes et(ou) sur certains produits logiciels universels, transférables et disponibles sur le marché, dont le système de gestion de base de données ORACLE avec l'interface standardisée du SQL. La plupart des autres propositions de la liste ci-dessus pourraient toujours être pertinentes mais elles devraient être interprétées de nouveau dans une certaine mesure.

(6) ARCHITECTURE FUTURE DES SYSTÈMES D'INFORMATION STATISTIQUE

(6.1) LE BUREAU DE STATISTIQUES RÉPOND-IL ENCORE À UN BESOIN?

Au chapitre 3 du présent document, on a fait remarquer que le progrès technologique a supprimé l'une des restrictions relatives à la production statistique : la nécessité de partager des ressources informatiques et para-informatiques rares et coûteuses. Ceci a enclenché un processus de décentralisation. Y a-t-il des limites naturelles à cette évolution ou la décentralisation cessera-t-elle seulement lorsque le bureau de statistiques se sera décomposé en plusieurs enquêtes statistiques distinctes? Autrement dit, l'avenir réserve-t-il au bureau de statistiques un rôle en tant qu'entité organisationnelle solide et autonome, ou les enquêtes statistiques seront-elles prises en charge par d'autres agences gouvernementales?

Personnellement, je crois que les bureaux de statistiques ont un rôle important à jouer dans la société, indépendamment des possibilités de décentralisation qu'offre à l'heure actuelle le progrès technologique; toutefois, je pense que nous devons nous poser les questions cruciales énoncées ci-dessus, car d'autres le feront.

En Suède, la production statistique a été centralisée sous sa forme actuelle au début des années 1960. Le besoin de rationaliser efficacement au moyen d'une technologie informatique centralisée constituait à l'époque la principale raison de la centralisation. Mais il y avait d'autres raisons tout aussi valables. D'une part, on croyait que seul un bureau de statistiques central solide pouvait se permettre de maintenir un développement méthodologique puissant de haute qualité et suffisamment important pour former une "masse critique". D'autre part, il y avait la nécessité de coordonner et d'intégrer les enquêtes individuelles dans les systèmes d'information statistique, d'après les modèles conceptuels unifiés tels que le système de comptes nationaux ainsi que les modèles socio-démographiques et socio-économiques.

Jusqu'à très récemment, les arguments technologiques à l'appui de la production statistique centralisée étaient acceptés de façon si générale que nous n'avons pas eu à recourir à d'autres arguments plus subtils. En conséquence, nous n'avons peut-être pas été aussi actifs que nous aurions dû l'être dans les domaines de la méthodologie et de l'intégration.

(6.2) BESOINS DES UTILISATEURS

Les besoins en matière de coordination et d'intégration reposent essentiellement et directement sur les besoins impérieux des utilisateurs. Par exemple, les clients de Statistics Sweden sont contrariés, et à juste titre, lorsqu'ils ont de la difficulté à localiser et à interpréter les données statistiques qu'ils recherchent, et lorsqu'ils doivent se rendre à plusieurs endroits à l'intérieur de l'organisation pour obtenir toutes les données dont ils ont besoin, y compris des conseils sur la façon de combiner les données provenant de ces différentes sources.

Qui plus est, un nombre croissant d'utilisateurs de statistiques expérimentés aimeraient mettre à l'essai des modèles et des hypothèses plus ou moins évolués avec des données statistiques officielles et parfois les combiner avec leurs propres données. Grâce au progrès technologique, ces utilisateurs conscients des possibilités qu'offre la technologie moderne auront toujours accès à leurs propres ordinateurs puissants. S'ils ne sont pas bien servis par le bureau de statistiques, ils feront valoir tous leurs droits afin d'obtenir de celui-ci des données statistiques brutes qu'ils utiliseront avec leurs propres données, logiciels et modèles dans les laboratoires de données qu'ils construisent et exploitent de façon indépendante. Si le bureau de statistiques veut pouvoir affronter cette concurrence, il doit se montrer dynamique, imaginatif et souple, et utiliser ses avantages relatifs au niveau de la compétence méthodologique et des possibilités de coordination et d'intégration.

(6.3) FAUT-IL Y REPENSER?

L'enquête statistique représente la pierre angulaire de nombreuses organisations statistiques. J'ai fait observer que la décentralisation suivie contribuera à renforcer davantage le pouvoir et le contrôle exercés par les enquêtes individuelles. Au niveau de la gestion, ce développement est positif à plusieurs égards. Il précise les responsabilités au sein de l'organisation et si quelque chose ne va pas, le responsable d'une enquête statistique peut difficilement blâmer tout le monde.

Par ailleurs, il y a les besoins des utilisateurs dont nous avons discuté à la section précédente et qui nécessitent d'autres solutions organisationnelles. Pour faciliter la localisation et l'interprétation des données, toutes les données statistiques d'importance doivent être bien documentées et l'être uniformément d'une enquête à l'autre. Ainsi, toutes les données statistiques doivent être accompagnées des métadonnées appropriées. Les métadonnées doivent être informatisées, et comme les données elles-mêmes, elles doivent être organisées selon les principes de bases de données uniformes. Bon nombre de bureaux de statistiques tentent depuis longtemps d'appliquer ces idées de différentes façons mais les résultats ne sont pas toujours encourageants. Dans le contexte de la concurrence actuelle, l'engagement dans ce travail ne sera pas suffisant pour certaines enquêtes et certains services. Plutôt, la mise en application systématique à travers l'organisation peut s'avérer une question d'importance capitale pour le bureau de statistiques en tant qu'entité autonome.

Les bases de données et les métadonnées ne suffiront pas aux besoins des utilisateurs. Ces données et métadonnées étant facilement accessibles, il est certain qu'ils combineront les données des différentes sources, à savoir les données provenant de différentes enquêtes et autres sources telles que les registres administratifs. Et, ils feront ces combinaisons, que nous les approuvions ou non. Traditionnellement, les bureaux de statistiques pouvaient se réfugier derrière leurs publications. L'enquête statistique donne lieu à des rapports et à des publications que l'utilisateur emploie à son gré en combinant les données qu'ils contiennent avec celles d'autres publications. Cette stratégie ne pourra être maintenue dans l'ère de la nouvelle technologie.

Il s'agit d'un problème complexe pour lequel il n'existe pas de solution simple. Toutefois, il semble évident que les bureaux de statistiques doivent prendre des initiatives dans le domaine des concepts et des classifications standardisés, un domaine dont l'importance n'est pas toujours pleinement reconnue par les responsables des enquêtes individuelles ni parfois par

les gestionnaires à des niveaux supérieurs. Par contre, ce domaine semble incontesté en ce sens que la plupart des personnes à l'extérieur d'un bureau de statistiques acceptent la responsabilité du bureau et reconnaissent sa compétence dans le domaine.

Les concepts et les classifications unifiés forment une excellente base pour la combinaison de données et la construction de modèles. Toutefois, il y a toujours des cas où la standardisation complète est impossible, et il importe que le bureau de statistiques joue également un rôle positif dans de telles situations, même s'il ne peut assumer la responsabilité des difficultés. Par exemple, il n'est pas possible de concilier les utilités différentes du registre administratif et d'une enquête statistique avec toutes les définitions. Néanmoins, les experts des bureaux de statistiques devraient se charger de trouver des moyens de contourner les problèmes afin d'exploiter de façon positive les capacités de la méthode statistique.

A l'aide de ces exemples, j'ai démontré qu'il existe un besoin global d'action et de contrôle administratifs en ce qui a trait aux enquêtes individuelles. On pourrait aller plus loin et dire que les nouveaux problèmes et possibilités nécessitent un nouveau concept d'enquête. Traditionnellement, l'enquête était modelée en une série d'étapes de production en commençant par le plan général et la collecte de données et en se terminant par les tables, rapports et analyses. La technologie moderne provoque entre autres l'affaiblissement matériel, logique et temporel des liens entre les données d'entrée et les résultats. Les résultats statistiques et les utilisations types de statistiques reposeront sur les combinaisons de données d'entrée provenant de nombreuses sources, et les données recueillies dans le cadre d'une enquête statistique seront utilisés à différentes fins par différents utilisateurs et à différents moments.

Par conséquent, si nous considérons l'enquête statistique comme la pierre angulaire de l'organisation statistique et du système d'information statistique, il serait peut-être plus adéquat d'envisager trois différents types d'enquêtes:

- enquêtes conçues en fonction de l'introduction des données; elles consistent à recueillir et à vérifier des données, à accomplir certaines tabulations et analyses courantes et à préparer les données en vue d'une utilisation ultérieure en les stockant, dans des bases communes de données, avec les métadonnées qui les accompagnent;

- bases communes de données; elles consistent à se charger des données de différentes enquêtes conçues en fonction de l'introduction des données et à former la base des enquêtes conçues en fonction de la sortie;

- enquêtes conçues en fonction de la sortie; elles consistent à utiliser les données existantes dans les bases communes de données et dans d'autres sources, à l'intérieur et à l'extérieur du bureau de statistiques.

L'UTILISATION DE NOUVELLES TECHNOLOGIES À STATISTIQUE CANADA

J. E. DOUCET,

Statistique Canada

L'UTILISATION DE NOUVELLES TECHNOLOGIES A STATISTIQUE CANADA

J.E. Doucet

14 février 1989

Direction de l'informatique
Statistique Canada
Ottawa K1A 0T6

(Original in English)

SOMMAIRE

Ce document donne un aperçu des nouvelles techniques que Statistique Canada est ou sera bientôt en mesure de mettre en oeuvre. Il décrit des applications intéressantes de nouvelles techniques auxquelles le Bureau a déjà recours et présente des hypothèses quant à l'orientation future de ces utilisations inédites. Finalement, l'auteur passe en revue des nouvelles techniques qui n'ont pas encore été mises à profit à Statistique Canada et s'interroge sur la possibilité de les incorporer à la panoplie des moyens technologiques en usage au Bureau.

INTRODUCTION

Depuis une vingtaine d'années, tous ceux qui, comme l'auteur, travaillent dans les bureaux centraux de la statistique, ont vu déferler plusieurs vagues informatiques. Chacune a permis d'augmenter sensiblement la puissance et la commodité de l'ordinateur tout en réduisant proportionnellement le coût de l'unité de traitement. Ces pas de géants nous ont encouragés à informatiser nos activités, à telle enseigne que l'ordinateur occupe aujourd'hui une place considérable. C'est ainsi que Statistique Canada a réussi à réduire, sans heurt ose-t-on croire, à quelque 4 400 un personnel qui atteignait les 6 000 à la fin des années 1960. Et pourtant, nous sommes plus productifs que jamais.

Mais l'informatisation ne s'est pas toujours faite en douceur. Nous avons dû traverser des étapes difficiles d'adaptation et de conversion, d'un système à l'autre; certains d'entre nous ont été appelés à apprendre plusieurs modes d'utilisation plus complexes les uns que les autres, dans le cadre de leurs fonctions; mais, en fin de compte, cela nous a permis d'aiguiser nos compétences, à force de résoudre toujours les mêmes problèmes apportés par chaque nouvelle vague.

Maintenant, les gestionnaires commencent à voir les nouvelles techniques d'un point de vue plus stratégique. Tirant profit des leçons du passé, ils sont plus critiques dans le choix de nouveaux moyens techniques, c'est-à-dire:

- qu'ils exigent que ces moyens s'intègrent facilement au dispositif informatique en place dans leur organisation et aux services offerts par cette dernière; et

- qu'ils tiennent à ce que les nouveaux moyens soient conformes aux diverses normes internationales qui se sont imposées ces dernières années.

C'est la situation qui règne aujourd'hui à Statistique Canada.

DISPOSITIF LOCAL D'ÉVALUATION DES NOUVELLES TECHNIQUES

Lors d'une réunion de planification organisée récemment à Statistique Canada, on a proposé le principe du cadre informatique intégré (CII); le CII est un univers offrant des services informatiques stables, uniformes et simples, dans un contexte relativement ininterrompu composé d'un matériel très diversifié. Une particularité importante du CII est qu'il permet d'accéder à n'importe quel service informatique depuis n'importe quel terminal ou poste de travail.

Un tel environnement est évidemment souhaitable parce qu'il fournit aux utilisateurs de services informatiques un ensemble uniforme de moyens d'intervention, contrairement aux systèmes classiques qui sont chacun dotés de leur propre protocole de dialogue. Mais il y a plus:

- les activités de soutien du système peuvent être concentrées dans des progiciels plus compacts à fonctions intégrées, ce qui simplifie la tâche de l'utilisateur;

- en réponse aux besoins d'exploitation, à des écarts de débit ou pour d'autres raisons, les applications peuvent passer d'un système à un autre;

- d'importantes économies de matériel deviennent possibles, grâce à l'interchangeabilité presque totale des applications; les soumissions d'achat peuvent être plus concurrentielles;

- on peut obtenir des remises quantitatives importantes sur l'achat de logiciel, étant donné que l'on achète davantage d'exemplaires d'un nombre plus restreint de programmes.

Le cadre informatique intégré présente un paradoxe intéressant. En informatique, on a toujours cru que stabilité et intégration étaient incompatibles avec innovation. Or, voici qu'aujourd'hui on peut prétendre que plus le cadre est stable et intégré, plus il est capable de réagir aux innovations technologiques et plus il lui est facile d'intégrer ces innovations à l'ensemble de ses services. Comment cela se fait-il? C'est simple; par définition, le CII fonctionne selon des normes industrielles internationales : SQL pour la gestion des données; langage C pour les compilateurs de troisième génération; modèle de l'ISO (interconnexion des systèmes ouverts) pour la communication interprocessus; UNIX et un petit nombre d'autres systèmes d'exploitation; etc. De plus en plus, l'innovation technologique qui aspire à la commercialisation suit ces sentiers officiels ou reconnus comme tels par la majorité. C'est dire que le CII a toutes les chances de bénéficier des progrès techniques présents et à venir, puisqu'il se place en plein centre de la tendance générale.

La transition au CII à Statistique Canada

A Statistique Canada, les principaux éléments d'un CII ont été mis en place au cours des deux dernières années.

- Dans le domaine crucial de la gestion des données, on s'est engagé à adopter les protocoles relationnels (notamment SQL) qui deviendront sous peu la norme au gouvernement fédéral, en matière de logiciel de gestion des données. Il y a plus de deux ans, le Bureau a choisi le logiciel ORACLE, qui peut fonctionner avec tous les systèmes de l'organisation.

- Lentement mais sûrement, on réduit le nombre de systèmes d'exploitation en usage. On a réussi à mettre un frein à la multiplication des systèmes d'exploitation de mini-ordinateurs. A Statistique Canada, les logiciels d'exploitation qui semblent se démarquer des autres sont le MVS, le DOS et UNIX; pour ce qui est de l'OS/2, il est probable qu'on le confinera aux serveurs de fichiers des réseaux locaux, du moins à court terme. Enfin, là où on utilise encore des systèmes d'exploitation propres à un constructeur, on essaie autant que possible de laisser de côté ou d'éliminer leurs particularités les plus distinctives.

- On procède doucement à la mise en place d'une infrastructure normalisée de communications fondée sur Ethernet, TCP/IP et SNA et on espère réaliser un jour l'ISO. L'interconnexion des systèmes se fait de plus en plus au moyen de passerelles d'Ethernet, de sorte que l'on abandonne progressivement les protocoles bilatéraux et les progiciels d'émulation.

- Le langage de troisième génération choisi d'emblée est le C, qui peut être utilisé avec tous les processeurs. Pour les nouvelles applications, il tend à remplacer le PL/I et il est largement utilisé pour la mise au point de systèmes au moyen de micro-ordinateurs.

- De plus en plus, les logiciels statistiques de fabricants ou élaborés dans les bureaux sont à fonctions intégrées et multicompatibles; par exemple, on peut se procurer le SAS pour trois types d'unités centrales et le nouveau progiciel de traitement des données d'enquêtes mis au point dans des bureaux locaux sera entièrement transférable.

Le CII est en pleine évolution mais il doit cohabiter avec des éléments de l'ancien univers et parfois s'y superposer; on comprend que, dans ces conditions, l'intégration ne soit pas aisée. Patience est le mot d'ordre puisqu'il en sera encore ainsi pendant 15 ans. Heureusement, les services les plus anciens, qui ne sont pas intégrés, seront graduellement écartés et éliminés, au profit du CII[1].

---

[1] Il est possible, dans certains cas, d'adapter des dispositifs anciens au CII; par exemple, il pourrait être possible de modifier le logiciel de traitement de données RAPID jusqu'au point où on pourrait lui adjoindre une interface SQL. On disposerait ainsi d'une partie des capacités du logiciel de récupération conçu par des sociétés indépendantes et fondé sur le SQL, pour des utilisations connexes au recensement.

Le CII dans la pratique

C'est dans cette optique de traitement intégré qu'il faut évaluer une nouvelle technique; si elle ne se conforme pas au modèle, deux solutions sont possibles:

- ne pas l'adopter; ou
- voir si le modèle pourrait être modifié de manière à l'accepter.

Les expédients n'ont pas leur place dans un cadre intégré; on ne doit jamais acquérir de nouveaux moyens techniques sous prétexte qu'ils satisfont à des besoins locaux concernant une application unique. La nouvelle acquisition doit s'adapter parfaitement à l'architecture globale ou être modifiée pour s'y plier.

En utilisant le modèle du CII, il devient possible d'évaluer les dernières technologies de l'heure.

SQL ET LES TECHNOLOGIES RELATIONNELLES

Parmi les nouvelles techniques, SQL est probablement la plus importante parce qu'elle permet d'intégrer des applications qui débordent le cadre de la machine de gestion de la base de données. Les distributeurs de logiciels privés considèrent SQL comme l'interface normale entre leurs programmes et les autres produits utilisables. Par exemple, SAS Institute, société presque exclusivement axée sur le marché, a annoncé récemment que toute sa gamme de logiciels pourrait dorénavant fonctionner avec SQL; autrement dit, il devient possible d'adjoindre aux bases de données relationnelles des applications de logiciels SAS.

La démarche relationnelle intéresse depuis longtemps les services statistiques, compte tenu de la nature de leurs activités. Le GSFD de Statistique Canada constitue un exemple intéressant. Le GSFD est un vaste ensemble de stratégies techniques visant à redéployer les progiciels généraux de traitement des données d'enquêtes. Ce projet est fortement axé sur la méthode relationnelle, à un point tel que le traitement même des données d'enquêtes risque d'être revu.

Le GSFD: une démarche relationnelle

La figure 1 illustre le premier étage de la division des fonctions qui constituent le GSFD.

On compte cinq grandes fonctions toutes reliées par un système complet de gestion de base de données relationnelles. Il serait impensable de vouloir réaliser le GSFD en l'absence du logiciel de gestion des données relationnelles, dont la puissance et la simplicité sont garants de la réussite du système.
Il peut être utile de donner quelques explications concernant la fonction 5, qui est un outil systémique de conception. Au départ, chaque fonction est élaborée isolément; par exemple, la fonction d'apurement et redressement automatique accepte, de manière interactive, une formule similaire à celle que l'on utilise pour les équations linéaires. Un jour, la conception individuelle cédera sa place à la conception globale qui agira au niveau de toute l'enquête. On y reviendra plus loin, sous la rubrique des systèmes experts.

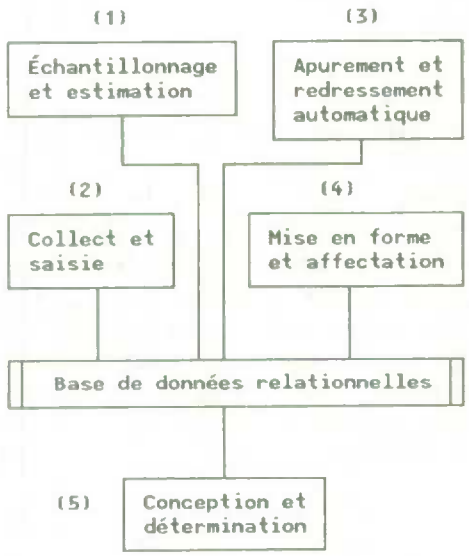


Schéma 1: Schéma des liens qui unissent les principaux composants du GSFD.

Grâce à la démarche relationnelle, le progiciel peut être axé sur la base de données. Le concepteur sera en mesure de déplacer les fonctions; par exemple, on pourra choisir de placer un ensemble de mise en forme plus tôt ou plus tard dans la séquence de traitement - qui sait, peut-être de la mise en forme préliminaire, dans les bureaux régionaux, à la remise en forme après affectation, sur le gros ordinateur? En outre, on pourrait commander l'affectation déterministe lors de la mise en forme préliminaire dans les bureaux régionaux, au lieu de plus tard dans le cycle de traitement, comme c'est le cas actuellement. Autrement dit, il devient possible de déplacer des fonctions selon les besoins particuliers en matière d'applications. Par conséquent, l'ancienne formule chronologique utilisée pour le traitement des données d'enquêtes n'est plus qu'un des multiples solutions possibles grâce au GSFD. En effet, le modèle pourra être organisé de plusieurs façons, pas nécessairement selon un ordre inhérent.

De manière plus générale, il devient possible de partager des programmes utilitaires avec d'autres fonctions. Par exemple, on procède au remaniement du programme de codage informatisé disponible sur l'ordinateur principal (ACTR - codage informatisé à reconnaissance de texte) pour l'intégrer au GSFD; il sera régi par le principe relationnel qui est utilisé dans le reste du système. Cela signifie que les grands avantages du codage informatisé seront accessibles à toutes les fonctions du GSFD. Le programme utilitaire qui fait le codage géographique des codes postaux subit le même sort; nul doute que d'autres suivront.

D'un point de vue plus technique, les versions à venir de SQL autoriseront l'adjonction au langage de fonctions propres à l'établissement. C'est ainsi qu'il sera possible de concevoir une fonction capable de calculer le séjour dans un établissement. Avec le temps, le nombre de tels algorithmes normalisés pourrait augmenter considérablement, à tel point que l'élaboration de systèmes pourrait un jour devenir rien de plus que l'assemblage d'une solution satisfaisante et

son adaptation à un système d'exploitation donné.

Le modèle relationnel et le nouveau matériel

L'importance du modèle relationnel est reconnue depuis longtemps - en principe - par Statistique Canada et d'autres instances. Cependant, ce n'est qu'au cours des dernières années que l'on s'est décidé à apporter au matériel les modifications nécessaires à la réalisation de logiciels de bases de données relationnelles. Ces logiciels doivent pouvoir traiter de très grands tableaux en mémoire et à très grande vitesse, ce qui nécessite des unités centrales rapides et beaucoup de mémoire bon marché. Nous assistons actuellement à l'arrivée de la première vague de systèmes informatiques dont l'architecture convient à l'exploitation de logiciels de gestion de bases de données relationnelles. Il y aura d'autres vagues qui viendront largement combler les vides dont souffre actuellement le logiciel relationnel, notamment en matière de processeurs en simultanéité, de processeurs vectoriels, de systèmes d'exploitation virtuels améliorés, de systèmes d'exploitation à 32 bits pour les grands micro-ordinateurs, de mémoires plus puissantes et moins chères, etc.

Le modèle relationnel et les bases de données réparties

Avec un SGBDR, les données procèdent d'un dépôt logique unique mais, physiquement, elles courent la chance d'être réparties entre plusieurs systèmes. L'état actuel d'évolution des logiciels de gestion de bases de données relationnelles (BDR) permet le rattachement des différentes tables d'une BDR à plusieurs unités centrales, sans que l'utilisateur ne s'en rende compte; le logiciel de gestion de la base de données, qui connaît le réseau, se charge d'assembler les données pour l'usager. Par contre, diverses parties de la base de données logiques peuvent être rattachées à des unités centrales différentes, mais non à l'insu de l'utilisateur. Au début, Statistique Canada a choisi cette dernière solution, parce que son réseau n'est pas encore (développé au maximum et qu'il est aux prises avec les problèmes particuliers qui découlent des impératifs de sécurité et de discrétion imposés par sa politique sur le réseau. Quoi qu'il en soit, la véritable base de données réparties demeure une solution techniquement possible (actuellement inexploitée) qui s'inscrirait dans l'esprit du CII actuel.

Il faut préciser que les "métadonnées" sont également implantées dans la base de données relationnelles. Ainsi, toutes les informations de conception et de détermination de la fonction 5 du GSFD peuvent être disposées sous forme tabulaire et gardées en accès direct au moyen de la même base de données logiques. Dans le cas de la fonction 2 (collecte et saisie), on peut avoir besoin de ce genre de renseignements pour la production de documents personnalisés de collecte de données; en fait, il s'agit d'une fonction répartie.

LES SYSTÈMES EXPERTS

Dans nombre d'organisations, on ressent le besoin profond de mettre à bon usage les systèmes experts (rétrogradés dernièrement du rang d'"intelligence artificielle"). A ce jour, l'engagement de Statistique Canada sur cette voie a été plutôt modeste, et pour cause. Certes, quelques services statistique ont réalisé des systèmes de démonstration intéressants, mais les grandes capacités de cette nouvelle technique sont encore largement inexploitées.

Cette année, Statistique Canada fera ses premiers pas dans ce domaine. Pour commencer, on s'attaquera probablement à la construction d'un ordinateur frontal destiné au SIGCA (Système itératif général de chaînage d'articles); dans ce cas, il s'agira vraiment d'une application "experte" pour laquelle la demande dépasse déjà l'offre. Un projet plus ambitieux, qui pourrait être réalisé en parallèle, concerne les problèmes de conception et de détermination des enquêtes, déjà soulevé au chapitre sur le GSFD.

Il importe d'intégrer les systèmes experts aux activités informatiques courantes. Voilà pourquoi nous entreprenons la mise au point de postes de travail Sun commandés par le logiciel UNIX, fondés sur le système PROLOG et jouissant de l'appui du logiciel ORACLE de gestion de bases de données relationnelles, présent ailleurs dans le CII. Ce projet est d'une grande importance car il permettra de faire profiter le secteur de la production des perfectionnements mis au point dans le milieu privilégié de la recherche-développement.

Les applications des systèmes experts sont généralement difficiles à délimiter, cachées comme elles le sont souvent au coeur même d'autres applications beaucoup plus vastes et aussi plus classiques. C'est pourquoi il faut limiter l'utilisation des systèmes experts aux applications qui s'inscrivent nettement dans le cadre global des services informatiques.

LES POSTES DE TRAVAIL

Statistique Canada ne se démarque pas par l'usage qu'il fait des postes de travail; on en compte quelque 1 400 qui sont régis par micro-ordinateurs. Comme c'est le cas ailleurs, on procède lentement au remplacement des unités par des appareils pilotés par le microprocesseur 80286 et d'autres encore plus puissants.

Au début, on trouvait des postes de travail isolés ou regroupés en petits réseaux locaux qui leur permettaient de partager les coûts de périphériques comme des unités de disques à grande capacité et des imprimantes à laser. Plus récemment, la puissance accrue des appareils et notre connaissance plus approfondie de leurs capacités nous ont incités à procéder à des opérations de traitement plus audacieuses. Cette optique nouvelle entraîne des changements auxquels nous assistons aujourd'hui:

- l'arrivée de puissants serveurs de fichiers pour les réseaux locaux;
- l'introduction d'outils de gestion de bases de données relationnelles adaptés aux réseaux locaux, qui viennent remplacer les systèmes à utilisateur unique ne fonctionnant pas avec SQL;

- le perfectionnement de protocoles de communications; du XNS (Xerox) on passe au TCP/IP (Internet), afin de faciliter les communications avec d'autres systèmes informatiques;

- l'intégration de systèmes à utilisateurs multiples pilotés par le 80386 (régis par UNIX, dans la majorité des cas) dans des réseaux locaux classiques (Ethernet).

Et tout cela se passe non pas au rythme de l'avance technologique mais bien en fonction de nos capacités financières.

Au Bureau, on utilise de manière intéressante les micro-ordinateurs. Par exemple, tout l'élément mise en forme et affectation du GSFD a été élaboré à partir de postes de travail de micro-ordinateur (d'abord avec le DOS, ensuite avec UNIX), au moyen du logiciel que l'on aurait utilisé si on s'était servi d'un ordinateur central. L'adaptation du logiciel a été moins ardue que prévue mais il ne faut toutefois pas penser que cela s'est fait tout seul. Autre avantage de ce genre de réalisations:leurs auteurs ont donné à d'autres intervenants (spécialiste du sujet et méthodologiste) l'occasion d'apporter le système dans leur bureau, à leur propre poste de travail, et de le mettre à l'essai comme bon leur semble. C'est là un excellent moyen de connaître les véritables besoins d'une application, dès le début.

MÉMOIRES AUXILIAIRES

Dans le domaine de la mémoire auxiliaire, le disque optique est sans doute la réalisation la plus intéressante.

La forme la plus courante du disque optique est le CD-ROM (disque compact à mémoire morte), qui sert principalement à la distribution de produits réalisés à partir de données CANSIM (Système canadien d'information socio-économique). Normalement, on recourt au CD-ROM lorsqu'il est nécessaire de distribuer un grand nombre de copies d'un produit mais que le temps compte peu. Autrement dit, le CD-ROM entraîne des coûts initiaux élevés et la production de l'original prend du temps.

Une variation moins connue de la même technologie est celle du WORM (write once read many), qui fait actuellement l'objet d'un projet de recherche à la Division du commerce international et à la sous-division de la recherche et des systèmes généraux de la Division des services et du développement informatiques. Le WORM se prête mieux à la distribution restreinte rattachée à des applications urgentes. Si on préfère, le WORM a des coûts de réalisation faibles mais son prix unitaire est élevé; il faut toutefois compter qu'il n'est pas nécessaire de produire un original, ce qui lui donne l'avantage dans les applications urgentes.

Dernièrement, on a réalisé en laboratoire des WORM pouvant être effacés[2]. Cette innovation signifie que le support optique pourrait entrer en concurrence directe avec le support magnétique classique; peut-être est-ce l'impulsion dont a besoin l'industrie pour adapter des mécanismes d'accès traditionnels à la technologie du disque optique. En effet, jusqu'à maintenant, la capacité d'entreposage des disques magnétiques s'est accrue à un rythme tel que seule l'arrivée du support optique effaçable pourrait prétendre les supplanter.

---

[2] Il faudra donc songer à modifier le nom de ce support avant de le lancer sur le marché.

Les appareils de lecture de WORM ne sont pas normalisés et ils ne s'inscrivent pas dans le CII, ce qui n'empêche que l'on pourrait quand même les exploiter, dans une certaine mesure:

- comme support local d'entreposage pour des répertoires, la Classification type des industries, la Classification internationale des maladies, etc.;
- comme moyen de distribution restreinte d'articles tirant normalement profit de la micrographie;
- comme système jetable de secours pour les micro-ordinateurs.

Lorsque leur utilisation se répandra et que la technique employée pour la gestion de leurs données se normalisera, il sera bon de réfléchir de nouveau sur la place qu'ils pourraient occuper dans le CII.

Au chapitre des supports classiques, Statistique Canada a parcouru le même chemin que la plupart des autres grandes organisations : densités accrues sur tous les supports; conditionnement amélioré (par ex., cartouches de bandes) et capacité totale en hausse constante. Il faudra évaluer plus à fond les possibilités d'investir dans les domaines de la robotique et du disque à composants solides.

IMAGERIE ÉLECTRONIQUE

Au Bureau, l'imagerie électronique se retrouve dans deux univers distincts. Dans le CII, des imprimantes à laser sont rattachées aux réseaux locaux, aux gros ordinateurs (IBM 3800) et aux systèmes ministériels (IBM 3820). D'un côté, ces systèmes d'impression ne cessent d'évoluer, de l'autre, les logiciels sont de plus en plus capables de régir les textes et les graphiques. Un bon exemple de cet état de choses est l'évolution du processeur de GML (langage général de mise en forme), dans les gros ordinateurs. Récemment, on a inauguré le traitement par formules[3] et de nouveaux repères ont été placés qui permettent les graphiques en direct et le téléchargement vers certaines imprimantes.

Le second univers du traitement des images, habituellement fondé sur la technologie Xerox, est assez différent de celui du CII. Il englobe le traitement d'images "sérieuses" destinées surtout à la photocomposition. Pour le moment, il n'existe aucune interface douce entre les deux univers. L'un des défis qu'il reste à relever dans le CII consiste à élaborer des liens commodes entre ces deux mondes distincts.

L'édition électronique de bureau est en quelque sorte un satellite de ces deux mondes. Actuellement, on pourrait comparer sa situation à celle du disque optique; il n'existe pas de

---

[3] Cette distribution khi carré est un exemple du traitement par formules rendu possible grâce au GML :

$$F(\chi^2) = \int_0^{\chi^2} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n-2}{2}} e^{-\frac{x}{2}} dx$$

normes spécifiques et son utilisation devrait demeurer marginale car cette technique ne s'intègre pas facilement.

LES PROGRÈS RÉALISÉS DANS LE MATÉRIEL

Comme on l'a précisé plus haut, bon nombre de progrès matériels semblent être rattachés de près aux nouvelles techniques relationnelles. Or, à maints égards, l'arrivée de logiciels de traitement relationnel peut être attribuée aux progrès réalisés dans le domaine du matériel :

- mémoires d'ordinateurs de plus en plus puissantes;
- unités centrales plus rapides;
- systèmes d'exploitation à mémoire virtuelle capables d'accéder à des
- espaces très vastes;
- disques à accès très rapide (y compris disques à composants solides);
- processeurs en simultanéité et processeurs vectoriels.

Il reste à savoir dans quelle mesure le logiciel non relationnel peut tirer profit de ces progrès. Chose certaine, le CII, le modèle à Statistique Canada, est bien placé pour en bénéficier.

En 1986, le Bureau a effectué une évaluation poussée des systèmes de traitement vectoriel. Pour procéder au test, on a créé une charge de travail représentative de traitement statistique que l'on a soumise à trois processeurs vectoriels :

- IBM 3090 - processeur capable de traitement vectoriel grâce à des instructions supplémentaires intégrées au matériel (les ordinateurs NAS procèdent de la même manière);
- CRAY XMP - ordinateur autonome à traitement vectoriel;
- AMDAHL VP-1200 - processeur vectoriel relié à l'ordinateur central AMDAHL de type scalaire.

Pour résumer, disons que les résultats ont fait miroiter des économies substantielles; malheureusement, les systèmes en question ne s'intègrent pas facilement au CII. Par exemple, la majorité d'entre eux utilisent encore le FORTRAN et ne peuvent fonctionner avec le SQL. Voilà pourquoi on a décidé d'attendre que les processeurs vectoriels se rapprochent davantage des grands courants informatiques; cette attitude est conforme à la démarche que nous avons choisie en matière de traitement en simultanéité. Après tout, il est de la plus haute importance que l'on parvienne à intégrer ces techniques aux opérations courantes. Pour certaines, il est encore trop tôt; mieux vaut attendre qu'elles se généralisent, si jamais elles prouvent leur viabilité commerciale.

RÉCAPITULATIF

Voilà un survol rapide des principales nouveautés technologiques auxquelles s'intéresse Statistique Canada. L'auteur a également voulu présenter le CII, qui servira de critère de base à toutes les évaluations locales de nouveautés techniques. Cependant, bien des domaines secondaires n'ont pas été touchés :

- techniques du CASE, qui commencent à peine à être intégrées aux gammes de produits des distributeurs de logiciels de bases de données relationnelles;
- les logiciels de traduction automatiques qui peuvent sans problème exister en périphérie du CII;
- les postes de travail analytiques tirant profit de graphiques interactifs évolués; etc.

L'auteur espère avoir éclairé le lecteur sur le mécanisme qui servira à évaluer toutes ces techniques, en vue de leur intégration. Des modèles comme le CII pourront certainement contribuer à nous écarter de voies sans issue où:

- nous n'avons pas deux systèmes identiques; et
- l'intégration est impossible.

L'UTILISATION DE LA TECHNOLOGIE À STATISTICS SWEDEN

L. LYBERG ET P. DEAN,

Statistics Sweden

INTRODUCTION

Statistics Sweden est le bureau national de la statistique de Suède. Cette agence produit la majorité des statistiques officielles de Suède. Statistics Sweden est chargé d'effectuer des sondages au nom des autorités centrales et locales et de fournir des services de consultation en matière de statistiques. L'agence peut également recevoir des mandats de particuliers, de compagnies et d'organismes.

Statistics Sweden compte quatre sections : statistiques du marché du travail, statistiques des zones, statistiques des entreprises et statistiques des particuliers. Ces services effectuent plus de 200 enquêtes par année. Les services statistiques, la collecte et le traitement des données, et la recherche et le développement relèvent des sections mentionnées ci-dessus. Statistics Sweden emploie actuellement 1 669 personnes, soit quelques centaines de moins qu'il y a dix ans.

Au début des années 1980, d'importantes compressions budgétaires ont amené un développement relativement rapide de ce que l'on pourrait appeler la technologie de pointe. De plus, il devenait de plus en plus difficile d'engager du personnel disposé à exécuter le travail manuel traditionnel d'enquête. Par exemple, Statistics Sweden avait l'habitude d'engager du personnel temporaire pour le traitement des données de recensement. En fin de compte, il est devenu impossible d'engager de la main-d'oeuvre temporaire en raison de la nature du travail et de la réglementation suédoise sur le travail et l'agence a dû s'en remettre à son personnel régulier pour les recensements. Ceci n'aurait pas été possible sans la technologie. La reconnaissance optique des caractères (ROC) ainsi que le codage automatisé constituent des exemples des nouvelles techniques d'abord introduites dans les méthodes de recensement. La ROC a remplacé la perforation à l'aide d'un clavier et le codage automatisé a assumé plus des deux tiers de la codification de profession.

Toute la nouvelle technologie n'est pas nécessairement attribuable aux compressions budgétaires ou à d'autres rationalisations. Les possibilités d'exécuter davantage de tâches ont également conduit à l'utilisation de la nouvelle technologie. C'est le cas de nos tentatives en vue de mettre au point un ordinateur portatif pour la collecte de données et de la technologie disponible pour diverses formes de statistiques sur l'utilisation des terres. Au cours des dernières années, Statistics Sweden a acheté plusieurs centaines d'OP et le paysage informatique a changé rapidement. On peut affirmer catégoriquement que le travail d'enquête diffère énormément par rapport à il y a cinq ou six ans.

Dans ce document, nous donnons quelques exemples ou plutôt un aperçu de l'utilisation de la technologie à Statistics Sweden. On y décrit d'abord le paysage informatique, puis nos tentatives en vue d'informatiser la collecte de données et le codage. Nous terminons avec des notes sur les aspects technologiques relatifs à la validation, au caractère privé et à la confidentialité, aux statistiques sur l'utilisation des terres et à la diffusion des données.

L'INFORMATIQUE À STATISTICS SWEDEN

Pour la plupart des gens, l'expression technologie de pointe se rattache aux ordinateurs. L'ordinateur symbolise la technologie de pointe dans une ère dominée et définie par cette même technologie. Le potentiel de n'importe quel bureau de statistiques est donc intimement lié à sa puissance informatique et à sa façon d'utiliser la puissance informatique dont il dispose.

Le parc informatique de Statistics Sweden est composé de deux processeurs centraux (un IBM 4381-3 et un NAS AS/8053), de quelques minis-ordinateurs (IBM S/36 et UNIX), et de 300 à 400 ordinateurs personnels. Quelque 50 OP sont utilisés dans le cadre de diverses expériences avec les réseaux d'OP.

Statistics Sweden possède deux emplacements et les processeurs centraux sont situées à Örebro (une ville située à environ 200 kilomètres à l'ouest de Stockholm) et communiquent avec les terminaux de Stockholm par des lignes téléphoniques louées à bail. Le IBM 4381-3 a une capacité de traitement de 2x2,5 millions d'instructions par seconde et une mémoire centrale de 32 millions d'octets. Le NAS AS/8053 a une capacité de traitement de 7 millions d'instructions par seconde et une mémoire centrale de 32 millions d'octets. Nos unités de mémoire comprennent un IBM MSS 3851 de 169 gigaoctets et une mémoire à disque d'une capacité totale d'environ 50 gigaoctets. Nous avons l'intention de remplacer nos unités de mémoire et nous envisageons un système d'archivage sur bande robotisé.

Ordinateurs personnels

Au moment de rédiger ce document, Statistics Sweden comptait entre 300 et 400 ordinateurs personnels et ce nombre devrait doubler dans un avenir très rapproché. A l'heure actuelle, on utilise principalement les OP dans le cadre de la bureautique, c'est-à-dire le traitement de texte pour des documents courants tels que la correspondance et les rapports, et même le traitement de texte à un niveau plus évolué, notamment l'édition électronique avec des tracés et des graphiques complexes. L'utilisation de tableaux électroniques tels que ceux du Lotus 1-2-3 pour l'élaboration de budgets et les calculs qui auparavant auraient nécessité une calculatrice entre également dans la bureautique. La majeure partie de l'analyse de données effectuée à Statistics Sweden se fait à l'aide d'OP SAS. On a déjà effectué la saisie des données sur de l'équipement spécial, soit CYBERDATA, mais cela aussi se fait sur OP maintenant.

La technologie actuelle permet de traiter de petites enquêtes en tout ou en partie sur un OP. Des études antérieures ont démontré que plus de 50 pour cent des produits statistiques de Statistics Sweden traitent moins de dix mille observations et peuvent être entièrement produits sur OP. A l'avenir, la principale tâche des processeurs centraux consistera donc à gérer les bases de données et à traiter les grandes enquêtes telles que le recensement de la population et du logement. Néanmoins, on prévoit effectuer certaines parties des grandes enquêtes avec des OP.

On retrouve trois types d'ordinateurs personnels à Statistics Sweden : le Macintosh de Apple, le "PC" d'IBM et ses compatibles et le PS/2 d'IBM. On a surtout recours au Macintosh pour le traitement de texte complexe (édition électronique) en raison de ses excellentes fonctions de rédaction de formules et de ses applications graphiques. Les compatibles "PC" d'IBM dominent toujours le marché et sont extrêmement concurrentiels au niveau des prix, du

nombre de fournisseurs et de la disponibilité des composants supplémentaires et des logiciels. On utilise les compatibles IBM pour le traitement de texte courant et le travail administratif. Le PS/2 est beaucoup moins répandu et offre certains avantages en ce sens qu'il peut fonctionner avec le système d'exploitation à disques (SED) et le nouveau système d'exploitation OS/2 d'IBM.

Les réseaux d'OP

Nous expérimentons actuellement divers réseaux et nous en examinerons quelques-uns dans le présent document. Le réseau est un ensemble de différents ordinateurs, terminaux et autres composants reliés ensemble de manière à pouvoir communiquer entre eux. Le réseau local d'OP diffère des autres réseaux du fait que la communication se fait à des vitesses très élevées à travers les câbles conçus pour les réseaux d'OP. Le réseau standard est constitué de serveurs, c'est-à-dire d'ordinateurs servis par des logiciels, des imprimantes, etc. et de postes de travail. De plus, certains réseaux sont dotés de systèmes de communication interréseau leur permettant de communiquer avec d'autres systèmes dont notamment un processeur central. Les passerelles, c'est-à-dire la connexion entre différents réseaux, peuvent également faire partie d'un réseau mais à l'heure actuelle, il n'existe pas de passerelles entre les réseaux de Statistics Sweden.

Quelques-uns des réseaux sont reliés au mini-ordinateur du Système 36 d'IBM. Il y a trois réseaux de Système 36 IBM pour un total de 90 noeuds, c'est-à-dire terminaux, imprimantes, etc. Le Système 36 est principalement utilisé en bureautique et ne convient pas à d'autres types de travail. IBM a cessé de produire ce système qu'elle remplacera par le AS/400. Statistics Sweden ne prévoit pas acheter de AS/400 mais a l'intention d'étendre son système de réseaux d'OP.

Il existe un réseau Macintosh qui utilise le logiciel de communication AppleTalk et est constitué de 20 noeuds : un serveur, quatre imprimantes et quinze postes de travail. Un autre réseau AppleTalk est composé de quatre postes de travail et d'une imprimante à laser.

On procède actuellement à un certain nombre de tests sur les "PC" d'IBM en utilisant l'anneau à jeton d'IBM et le Réseau local de communications d'IBM, comme logiciel. Pour faire partie de l'un de ces réseaux, l'OP doit être équipé d'un adapteur d'anneau à jeton. Ces ordinateurs sont reliés par un système de câbles IBM.

Logiciels

Statistics Sweden a élaboré la plupart des logiciels utilisés sur les processeurs centraux pour la production de statistiques. Il s'agit de programmes standard et généraux utilisés dans la plupart des enquêtes de Statistics Sweden pour compiler les tables (TAB68) et les graphiques (Cue-Chart), calculer les variances (SMED) et exécuter tout autre traitement sur ordinateur nécessaire à la production de statistiques descriptives. Comme les autres agences nationales, Statistics Sweden ne fait pas couramment d'analyses de données, bien que cette tâche puisse faire partie d'un travail de consultation et qu'elle soit exécutée dans certaines enquêtes continues. Nous avons également élaboré un logiciel de sélection d'échantillons adapté aux conditions suédoises (nos registres uniques, etc.)

Au début des années 1980 nous avons compilé les bases de données et créé AXIS, un logiciel de gestion de base de données pour la documentation, la présentation et la distribution de tables et de séries chronologiques. AXIS permet à l'utilisateur de générer des tables adaptées à ses besoins spécifiques. Il se distingue des autres gestionnaires de base de données en ce sens que les données sont stockées de façon à permettre la production de tables et de séries chronologiques uniques. AXIS est actuellement utilisé par le bureau central de la statistique du Danemark pour sa base de données de municipalités et à Genève par la Commission économique pour l'Europe.

Par ailleurs, le logiciel qu'utilise Statistics Sweden est un produit de l'industrie et est disponible sur le marché. Comme nous l'avons mentionné précédemment, nous réservons surtout les OP pour le traitement de texte, le travail administratif et un peu d'analyse de données. Nous avons essayé de limiter le nombre de logiciels différents et de promouvoir un assortiment standard de logiciels utilisés à l'agence. On a jugé qu'il n'était pas économique d'avoir trop de logiciels différents et se chevauchant, les responsables de l'installation, de l'entretien et de la consultation ne pouvant pas connaître à fond tout ce qu'il y a sur le marché. Ceux qui choisissent d'utiliser un logiciel "rare" doivent en assumer l'entière responsabilité et ne pas demander l'aide des informaticiens de l'agence. Cependant, on peut facilement obtenir de l'aide pour le logiciel approuvé par l'unité informatique, c'est-à-dire le SAS.

Politique

En gros, à Statistics Sweden, la politique en matière d'informatique veut que la technologie conduise éventuellement à une production de statistiques extrêmement décentralisée. Le personnel d'une enquête sera responsable de tous les aspects de l'enquête depuis la conception du questionnaire et la collecte de données jusqu'à la présentation et à la diffusion. Dans ce système décentralisé, le statisticien d'enquête se doit d'être un "homme-orchestre" en ce sens qu'il est responsable de l'enquête depuis le début jusqu'à la fin. Il est même censé produire ses propres publications à l'aide de logiciels de traitement de texte complexes (édition électronique). Cela exige beaucoup de la part du statisticien car il doit bien connaître toute une variété de logiciels et maîtriser tous les aspects de la méthode d'enquête. Il doit constituer une unité autonome appuyée par un assortiment bien choisi de progiciels généralisés. Ces progiciels sont ceux élaborés par l'industrie et par Statistics Sweden.

COLLECTE DE DONNÉES ASSISTÉE PAR ORDINATEUR

On travaille à l'élaboration de la méthode d'interviews téléphoniques assistées par ordinateur (ITAO) depuis près de vingt ans. La plupart des utilisateurs de l'ITAO rapportent de bons résultats. Néanmoins, les avantages présumés de l'ITAO en ce qui a trait à la qualité et aux coûts comparativement aux interviews téléphoniques réguliers semblent reposer sur des suppositions plutôt que sur des études méthodologiques (voir Nicholls et Groves, 1986, et Groves et Nicholls, 1986).

Ces dernières années, Statistics Sweden a travaillé à l'élaboration d'un dispositif encore plus perfectionné pour un différent type de collecte de données. Nous visons un dispositif général pour la collecte de données assistée par ordinateur (CDAO). Le défi consiste à trouver un dispositif qui pourrait être utilisé dans un cadre standard d'ITAO centralisé pour les interviews

téléphoniques décentralisées réalisées chez l'enquêteur, pour les interviews en tête à tête effectuées presque n'importe où et pour l'introduction de données sur les prix, l'abondance des récoltes, etc. Pour ce faire, nous avons besoin d'un ordinateur portatif.

Voici un résumé des raisons justifiant la mise au point d'un ordinateur portatif.

1. Statistics Sweden a besoin d'un système automatisé et décentralisé de collecte de données car nos enquêteurs sont dispersés à travers le pays. Nos ressources d'automatisation seraient limitées si l'on utilisait seulement la méthode ITAO, centralisée ou décentralisée.

2. Une partie du travail d'enquête exécuté à notre agence est effectuée en vertu de mandats. Pour conserver sa part du marché, Statistics Sweden doit être au premier rang en matière de technologie.

3. Le coût, la qualité et l'à-propos sont vitaux pour la production de statistiques. Tout dispositif permettant de réaliser des économies et d'améliorer la qualité devrait être mis à l'essai.

Les plans de la CDAO suédoise exposés en 1984 ont été présentés par Lyberg (1985). Ces plans comprenaient un ensemble de spécifications ayant trait à l'ordinateur comme tel (poids, ergonomie, etc.), à la fabrication de l'ordinateur, à l'élaboration du logiciel et aux essais. On a soumis ces exigences à plusieurs fabricants, mais un seul était disposé à relever le défi. La plupart des compagnies prétendaient qu'il serait trop difficile de répondre à nos exigences relativement au poids. (Pour permettre une collecte de données ergonomique, on avait décidé que l'ordinateur devait peser moins d'un kilo et que l'écran devrait permettre l'affichage de 24 caractères sur 80.)

On peut trouver ces exigences détaillées dans Danielsson et Marstad (1983). Brièvement, on exigeait (a) les facteurs ergonomiques mentionnés, (b) que l'ordinateur soit utilisé dans toutes sortes de milieux de collecte de données, (c) que les textes, messages et données soient transmis par réseau téléphonique, ce qui supposait une réduction considérable des données, (d) que le stockage se fasse dans une mémoire de 512 kilo-octets, (e) que l'ordinateur fonctionne sur batteries pendant une journée sans être relié au réseau, (f) que le temps de réponse n'excède pas deux secondes, (g) que les applications puissent être transférées d'un ordinateur central aux ordinateurs portatifs en moins de 48 heures, et (h) que le système puisse être agrandi et contrôlé.

Tests

Le projet a débuté au début des années 80 par l'étude de divers systèmes ITAO américains. Pour acquérir de l'expérience au niveau de la CDAO, on a effectué en 1982 un premier test d'introduction de données sur des ordinateurs portatifs. On a utilisé un ordinateur appelé MICRONIC pour recueillir les prix dans les magasins pour l'indice des prix à la consommation. Généralement parlant, les résultats de ce premier test étaient encourageants. Bien sûr, l'ordinateur mis à l'essai avait de nombreux défauts dont un petit écran et une capacité de mémoire générale assez réduite. Mais les enquêteurs appréciaient la nouvelle technologie et l'évaluation technique du test a fourni beaucoup d'indications précieuses pour les élaborations subséquentes.

En 1984, on a éprouvé la méthode ITAO dans un cadre normal pour les installations d'ITAO avec des terminaux à écran reliés à un processeur central. Comme nous savons que de telles mises en place fonctionnent, cette étude portait principalement sur les réactions des enquêteurs. Dans l'ensemble, elles étaient favorables.

Afin de tester les réactions des répondants aux micro-ordinateurs portatifs, on a effectué une étude en 1985 à l'aide d'un ordinateur portatif, Data General/One, qui était lourd mais disponible sur le marché. Il n'y a eu aucune réaction négative de la part des répondants.

En 1986, quelques prototypes portatifs ont été livrés. Ces prototypes comportaient des défauts considérables vu les exigences qui avaient été établies. Le défaut le plus remarquable était la pesanteur de l'ordinateur. Malgré tout, vers la fin de 1986, on a entrepris une série de tests sur le nouvel équipement. Ces tests se sont déroulés dans le cadre de l'enquête sur le choix des partis politiques et de l'enquête sur la population active.

En 1987, le projet a été grandement modifié. En effet, le fabricant suédois a admis qu'il était incapable de produire un appareil supérieur aux prototypes déjà élaborés. La compagnie a déclaré qu'elle ne possédait pas le savoir-faire ni les autres ressources nécessaires pour remplir toutes les exigences techniques fixées par Statistics Sweden. Même si nous attendions cette déclaration depuis un certain temps déjà, elle n'a pas manqué de nous décevoir considérablement. Le logiciel et les solutions techniques élaborés dans le cadre du projet étant en majeure partie indépendants du matériel, nous avons commencé à chercher un ordinateur déjà disponible sur le marché. Nous travaillons actuellement avec le Toshiba 1200.

Essai du Toshiba 1200 dans le cadre de l'enquête sur la population active

En 1988, le Toshiba 1200 portatif a été mis à l'essai dans le cadre de l'enquête sur la population active. Trente-cinq enquêteurs ont réalisé en tout 821 interviews et l'objet de l'étude consistait à éprouver le système de TED et le logiciel. Il s'agit de la première étape d'une étude de faisabilité de grande envergure qui comprendra également des comparaisons méthodologiques de la qualité et des coûts. Parmi les exemples de facteurs à l'étude pendant cette première étape, on retrouvait la fiabilité du système, la visibilité à l'écran, les qualités ergonomiques, la fiabilité de l'ordinateur portatif, la communication entre le processeur central et l'ordinateur portatif, et les réactions des enquêteurs et des répondants. Aucun des 35 ordinateurs Toshiba 1200 n'a présenté de problème majeur pendant le test. Le personnel du bureau central et les enquêteurs se tenaient au courant des problèmes en les consignant dans des registres. Ces registres comprenaient de l'information sur la communication entre les enquêteurs et le personnel du bureau central ainsi que des descriptions des divers problèmes survenus en cours d'interview.

Même si le système fonctionne bien, il requiert encore beaucoup d'améliorations. Évidemment, le problème du poids de l'ordinateur portatif n'a pas encore été résolu. Le poids actuel de 5,2 kilos est inacceptable pour le type d'appareil que nous visons. Nous avons également eu des problèmes avec les batteries. Dans certains cas, elles ne duraient pas aussi longtemps que prévu. La formation de deux jours pour les enquêteurs semblait trop courte; la communication intense entre les enquêteurs et le personnel du bureau central au cours des premières journées du test vient appuyer cette observation. Il est bien évident que les enquêteurs ont besoin de plus de temps pour s'entraîner avant d'entreprendre le véritable travail d'enquête. Le test a également fait ressortir un certain nombre de défauts reliés au logiciel, mais nous n'en parlerons pas maintenant.

On a demandé aux enquêteurs et à un échantillon de répondants de nous faire part de leurs réactions à l'égard de cette nouvelle technique. Les répondants étaient plus ou moins indifférents et certains enquêteurs n'étaient pas très emballés par l'ordinateur portatif. Ils se préoccupaient de l'efficacité de l'interview, du stress physique et du fait que l'interview se trouvait rigoureusement contrôlée. Toutefois, la plupart des enquêteurs trouvaient très intéressant de travailler avec l'ordinateur portatif.

Quelle sera la prochaine étape?

Les plans de fabrication d'un nouvel ordinateur portatif extrêmement léger sont actuellement en suspens. L'élaboration du logiciel se poursuit et ce dernier peut s'adapter à une installation ITAO récemment mise en place et comprenant huit postes de travail. Une série d'études en cours contribuera, nous l'espérons, à combler les lacunes dans nos connaissances sur les possibilités de réalisation des ordinateurs portatifs. En ce qui a trait aux ordinateurs portatifs comme tels, nous sommes limités au marché existant. On peut identifier différentes options selon l'élaboration des ordinateurs portatifs existants. Premièrement, on peut étendre les installations centralisées de l'ITAO au maximum grâce aux logiciels déjà disponibles. Deuxièmement, si de nouveaux ordinateurs portatifs font leur apparition sur le marché, Statistics Sweden pourra continuer à épurer son concept original. Troisièmement, si on ne peut se procurer que des ordinateurs lourds, on pourra augmenter l'ITAO centralisée de façon à englober l'ITAO décentralisée effectuée chez les enquêteurs. Quatrièmement, si, encore une fois, le marché n'offre que des ordinateurs lourds, notre concept pourrait englober l'ITAO et les interviews personnelles assistées par ordinateur mais exclure les types d'introduction de données où l'on doit transporter l'ordinateur pendant le processus de collecte de données. Pour l'instant, les solutions trois et quatre semblent les plus plausibles.

VALIDATION

La validation, c'est-à-dire le processus visant à assurer que les données recueillies remplissent certaines exigences de qualité, est reliée à une longue tradition à Statistics Sweden. La validation permet non seulement d'éliminer les valeurs erronées, mais elle sert également d'instrument de contrôle de qualité pour la collecte de données. Nous pouvons identifier deux formes principales de validation, la micro-validation et la macro-validation. En micro-validation, le contrôle des observations individuelles est effectué en suivant des règles spécifiques pour détecter les erreurs et corriger ou du moins régulariser les observations erronées. En macro-validation, on détecte les erreurs en contrôlant les groupes d'enregistrements.

La micro-validation permet d'enrayer certaines erreurs prévues attribuables à la négligence ou à la méprise du répondant. Les enquêtes devraient être conçues de façons à minimiser le risque de telles erreurs, ce qui signifie que dans le cas des enquêtes bien élaborées, ces taux d'erreurs sont plutôt bas et que la validation permet de corriger les erreurs au moment de la saisie des données. Si l'unité de saisie des données n'est pas assez puissante, on peut recourir à des logiciels de micro-validation généralisés. GRAN78, GRUS et FELOC en sont des exemples. Jusqu'à présent, nous en savons très peu sur l'effet de la validation sur la qualité. Certains indices permettent de croire que la micro-validation apporte très peu et que dans certains cas l'impact peut même est préjudiciable. Par conséquent, il est nécessaire de bien évaluer les pratiques courantes de micro-validation. Le recours à la micro-validation entraîne un risque

considérable de surestimation de la qualité des données. Il arrive souvent que lorsque les données ont subi plusieurs contrôles de micro-validation, ni les données ni le système de production d'enquête ne sont revus de façon critique.

La maintenance du logiciel présente certains problèmes. Les programmes sont désuets et limités. GRAN78 est rédigé en langage d'ASSEMBLAGE et GRUS, en langage APL. A Statistics Sweden, la compétence au niveau de ces langages de même que le besoin général de maintenir cette compétence à d'autres fins diminuent rapidement. L'ensemble actuel de programmes est sérieusement limité. Il est impossible de corriger simultanément deux ou plusieurs fichiers, par exemple, le fichier ménage et le fichier particulier dans une même enquête.

Statistics Sweden prend part au Projet de calcul statistique international. On a étudié des logiciels élaborés au Canada (CAN-EDIT), en Hongrie (AERO) et en Espagne (DIA) lesquels se sont avérés inadéquats pour les fins suédoises. Des études en cours se penchent sur le système BLAISE, élaboré aux Pays-Bas, et sur le système portatif canadien GEIS.

La macro-validation est plus rationnelle, mais ses résultats ne se manifestent pas immédiatement. Les erreurs détectées par la macro-validation peuvent être éliminées dans les enquêtes à venir, mais il n'est pas facile d'y remédier dans l'enquête en cours; par exemple, la macro-validation pourrait suggérer l'insertion d'une nouvelle procédure de redondance.

Le rôle de la validation devrait être étendu afin que celle-ci devienne un autre instrument de contrôle de qualité. Ainsi, elle serait essentiellement utilisée pour évaluer la qualité de l'enquête et pour fournir de l'information sur la qualité, et ce, jusqu'au niveau de la valeur de la variable simple. Il est tout particulièrement important que les valeurs des variables simples soient aussi précises que possible si l'on doit stocker les données dans une base de données. Le rôle de la validation ne devrait pas se résumer à la correction d'erreurs embarrassantes et à la mise au net des fichiers de données.

Les premières politiques de validation prévoyaient que le petit nombre d'anomalies et les valeurs manquantes ne perturberaient pas le traitement des données subséquent. Au United States Census Bureau (voir Ogus et al., 1967), on appliquait une politique selon laquelle cinq pour cent des données seulement pouvaient être corrigées (généralement par imputation) pendant le processus de validation. L'utilisation de méthodes évoluées d'imputation pour traiter les absences de réponses et le fait que la validation est perçue comme un instrument de contrôle de qualité font ressortir le besoin d'une nouvelle politique de validation.

Dans certaines applications de Statistics Sweden, la validation se fait de façon approfondie avec de nombreux contrôles; cela peut entraîner un grand nombre d'indications d'erreurs. Comme une fraction seulement de ces indications représente des erreurs véritables, les responsables des enquêtes qui ont souvent recours à la validation devraient penser à affecter davantage de ressources aux mesures préventives telles que les préenquêtes, l'élaboration de questionnaires améliorés et la formation des enquêteurs. Cependant, un risque de "suroptimisation" est inhérent à ce que nous appelons "la méthode du dispositif de contrôle". L'existence même d'un dispositif de contrôle semble justifier son utilisation. Le recours aux techniques disponibles, qu'elles soient nécessaires ou non, est répandu dans le travail relatif aux erreurs non attribuables à l'échantillonnage, un domaine où certains instruments ont été mis au point (notamment les contrôles de validation et de perforation) et où il reste à développer d'autres outils nécessaires (comme les instruments de contrôle pour certaines erreurs de mesure). Néanmoins, il n'est pas nécessaire de contrôler tout ce qui peut l'être. Comme l'ont

déclaré les Nations Unies (1982) : "L'expansion progressive des procédures de validation et d'imputation informatiques a provoqué un malaise chez certains utilisateurs et producteurs de données. Après tout, la collecte de données vise à découvrir ou à confirmer certains éléments de vérité sur la population. Apporter des modifications considérables aux données recueillies avant de rendre celles-ci disponibles pour l'analyse constitue une violation d'un principe fondamental en collecte de données, à savoir le respect de l'intégrité des données."

On doit donc pondérer soigneusement la validation par rapport à d'autres mesures de contrôle. Actuellement, Statistics Sweden estime que l'on consacre trop de ressources à la validation. Celle-ci devrait être considérée comme un problème méthodologique relatif à l'enquête plutôt que comme un problème informatique. Sans diminuer la qualité des données, il est possible de réduire les coûts de validation en choisissant soigneusement les contrôles de validation, en effectuant la majeure partie de la validation à l'étape de saisie des données et en utilisant la macro-validation plutôt que la micro-validation.

Un nouveau système de validation facile à utiliser et reposant sur ces principes est présentement à l'étude. On a élaboré et distribué un premier plan de ce système. Les réactions varient. La validation a toujours été un sujet à controverse à Statistics Sweden. Toutefois, on a décidé de poursuivre le développement. On devrait tenir compte des autres travaux du même genre qui sont effectués à l'agence, y compris la validation en tant que partie intégrante du système de collecte de données assistée par ordinateur et l'intégration de la saisie des données et de la validation.

CODAGE AUTOMATISÉ

Le codage constitue une opération importante dans le recensement de la population et des logements suédois ainsi que dans d'autres enquêtes effectuées par Statistics Sweden. Le codage transforme l'information verbale en numéros de code afin de faciliter le traitement des données. Effectuée à la main, cette opération comporte un risque d'erreurs. Plusieurs études révèlent que le taux d'erreurs peut être assez important lors du codage de variables complexes telles que la profession et la branche d'activité. Une étude portant sur l'évaluation du codage dans le recensement de 1970 en Suède a produit les résultats suivants.
Le tableau indique que les variables à plusieurs chiffres sont difficiles à coder. Mais les variables à un chiffre, considérées au départ comme faciles à coder, sont elles aussi très souvent mal codées.

Lors de la préparation du traitement des données du recensement de 1975 en Suède, on a tenu compte de nouvelles procédures de contrôle de qualité ainsi que de changements administratifs. Ceci a donné lieu à un programme de vérification renforcé et les taux d'erreurs ont diminué pour toutes les variables. Une étude d'évaluation a fourni les résultats suivants.

On retrouve dans Lyberg (1981, 1983) des révisions approfondies des études suédoises sur les taux d'erreurs au niveau du codage.

Dans les années 1970, il semblait important d'analyser en détail le codage manuel pour examiner les possibilités d'informatisation. Il avait été démontré que de meilleures techniques de contrôle manuel augmenteraient l'efficacité du codage manuel, mais cette amélioration n'était pas suffisante. L'automatisation s'imposait. A l'époque, le U.S. Census Bureau envisa-

Tableau 1. Estimation des taux d'erreurs dans le codage de l'activité économique lors du recensement de la population suédoise en 1970

Variable	Code	Taux d'erreurs exprimé en pourcentage (population totale)
1. Lien avec le chef du ménage	1 chiffre	4,3
2. Type d'activite	1 chiffre	4,7
3. Profession	3 chiffres	13,5
4. Situation de famille	1 chiffre	3,7
5. Branche d'activité	4 chiffres	9,9
6. Lieu de travail	1 chiffre	8,9
7. Moyen de transport utilisé pour se rendre au travail	1 chiffre	11,5
8. Nombre d'heures travaillées	1 chiffre	4,4

Tableau 2. Estimation des taux d'erreurs dans le codage de l'activité économique lors du recensement de 1975 en Suède

Variable	Code	Taux d'erreurs exprimé en pourcentage (population totale)
1	1 chiffre	0,6
2	1 chiffre	0,6
3	3 chiffres	7,8
4	1 chiffre	0,5
5	4 chiffres	3,5
6	1 chiffre	1,0
7	1 chiffre	0,5

geait le codage automatisé depuis un certain temps déjà (voir O'Reagan, 1972) mais la méthode n'avait jamais été appliquée à de vraies enquêtes.

Le codage automatisé comporte quatre opérations. Premièrement, il faut constituer un dictionnaire stocké en mémoire. Deuxièmement, les descriptions verbales sont introduites dans l'ordinateur. Troisièmement, les descriptions verbales sont assorties au dictionnaire et des codes sont assignés dans les cas de correspondance. Quatrièmement, on doit effectuer une évaluation régulière du processus automatisé de manière à contrôler le coût et la qualité.

Plusieurs problèmes méthodologiques sont inhérents au codage automatisé. Les plus intéressants sont ceux associés au dictionnaire et aux algorithmes correspondants. Il est préférable que le dictionnaire lui-même soit constitué par un programme machine. Toutefois, les données à traiter doivent consister dans des descriptions manuellement codées par des experts et de préférence, sélectionnées dans la première partie de l'enquête à coder. Lors de la constitution du dictionnaire, on utilise deux listes de descriptions codées par des experts. La première liste est composée de descriptions codées par numéro de code. Le même fichier, où les descriptions sont codées par ordre alphabétique, forme la deuxième liste. On choisit les descriptions les plus fréquentes de la première liste pour les inclure dans le dictionnaire. La deuxième liste est utilisée pour vérifier si on a assigné le même numéro de code à ces descriptions fréquentes. Il arrive souvent qu'une description verbale spécifique puisse être associée à différents numéros de codes et la deuxième liste aide à identifier ces descriptions ambiguës de façon qu'elles puissent être supprimées du dictionnaire. L'algorithme correspondant est très simple. Seules les correspondances exactes sont utilisées, ce qui signifie qu'une certaine proportion des descriptions verbales sont renvoyées au codage manuel lorsque le codage automatisé est appliqué. Le U.S. Census Bureau a concentré ses recherches sur des algorithmes correspondants plus complexes en utilisant la pondération. Ce travail est expliqué dans Appel et Scopp (1985) et Knaus (1987). Statistics Sweden n'a pas encore utilisé ni même expérimenté ces algorithmes.

Les applications du codage automatisé par Statistics Sweden comprennent le codage des biens dans les enquêtes sur les dépenses de ménage (EDM) de 1978, 1985 et 1988. Le plan de l'EDM de 1978 comprenait la livraison continue d'agendas tenus par les répondants pendant l'année de l'enquête, ce qui fait que les éléments étaient traités par cycle. En 1978, le dictionnaire initial reposait sur un codage expérimental effectué l'année précédente. Au cours de l'année de l'enquête, le dictionnaire a fait l'objet de 16 révisions, passant de 1 459 descriptions à 4 230. Le degré de réussite peut être évalué par la proportion de descriptions codées par l'ordinateur (degré de codage). Le tableau 3 indique le volume du dictionnaire ainsi que le degré de codage pour les 33 cycles.

Le degré de codage pour l'ensemble des cyles s'élevait à 65 pour cent. On a utilisé deux différentes formules d'agenda, ce qui explique les diminutions brusques dans le degré de codage. L'une des formules était beaucoup mieux adaptée au codage automatisé que l'autre.

On a utilisé essentiellement les mêmes procédures pour l'enquête sur les dépenses de ménage de 1985. Le tableau 4 affiche les données des 17 premiers cycles de cette enquête.
L'EDM de 1988 est toujours en cours, mais nous savons que le degré de codage se maintient autour de 80%

Voici quelques remarques sur le codage automatisé des enquêtes suédoises sur les dépenses de ménage. (1) Des études d'évaluation spéciales révèlent que le taux d'erreurs pour la partie codée automatiquement est d'environ 1%. (2) Le codage automatisé est légèrement moins coûteux (2 à 5%) que le codage manuel. (3) L'automatisation a engendré un meilleur contexte général de codage en ce sens que le codage manuel qui reste est devenu plus intéressant pour

Tableau 3. Volume du dictionnaire et degré de codage pour les 33 cycles de l'EDM de 1978

Version du dictionnaire	Cycle	Nombre de descriptions au dictionnaire	Degré de codage (%)
1	1	1 459	56
2	2	1 554	63
3	3	1 760	67
4	4	2 228	66
5	5,6,7	2 464	68,68,63
6	8	1 632	64
7	9	1 990	53
8	10,11	2 451	69,66
9	12	2 866	61
10	13	3 065	68
11	14	3 613	58
12	15,16	3 752	72,73
13	17,18	3 832	39,70
14	19,20	4 011	65,73
15	21,22	4 229	51,72
16	23,24,25, 26,27,28, 29,30	4 230	64,67,62,67 72,65,67,50
17	31,32,33	4 230	65,39,67

Tableau 4. Volume du dictionnaire et degré de codage pour les 17 premiers cycles de l'EDM de 1985

Version du dictionnaire	Cycle	Nombre de descriptions au dictionnaire	Degré de codage (%)
1	1,2	1 985	81,78
2	3	2 029	82
3	4,5	2 063	82,81
4	6	2 126	82
5	7	2 156	83
6	8	2 176	82
7	9	2 207	81
8	10	2 228	83
9	11,12,13	2 272	83,83,82
10	14,15,16	2 370	83,80,81
11	17	2 446	83

les codeurs et les tâches administratives ont été réduites. (4) Il ne semble pas utile de faire des révisions approfondies après un certain point. On obtient bientôt un degré de codage stable et même une importante mise à jour du dictionnaire ne peut accroître le degré de codage de

façon considérable. En admettant que la première version du dictionnaire ait été constituée conformément aux méthodes que nous avons décrites, le degré de codage automatisé est très peu coûteux. Il coûte moins de 50 $ pour des dizaines de milliers de descriptions.

les logements suédois constitue une autre importante application. Les descriptions verbales relatives à la profession sont codées et mises en forme. Pendant le processus de mise en forme, des signes et préfixes spéciaux sont retirés et les parties restantes de la description sont introduites en mémoire en une série. Chaque enregistrement est assorti à un fichier contenant de l'information sur la population économiquement active. Il en résulte un fichier assorti contenant la description verbale mise en forme ainsi que de l'information auxiliaire pour chaque enregistrement. Le fichier mis en forme est assorti à un dictionnaire stocké en mémoire et un numéro de code est assigné chaque fois que l'on trouve une description de profession dans le dictionnaire. Les enregistrements contenant des descriptions qui ne figurent pas dans le dictionnaire sont codés manuellement.

Le codage manuel est assisté par ordinateur et effectué en deux étapes sur des consoles de visualisation. Le codage manuel initial se fait sans accès aux questionnaires et rapidement. Les enregistrements qui ne peuvent être codés sont laissés "en blanc" et traités lors du deuxième codage manuel où le codeur a accès aux questionnaires.

Le dictionnaire se divise en deux parties : un dictionniare principal, PLEX, et un secondaire, SLEX. Le PLEX contient des descriptions sans équivoque et est scruté en premier. Le SLEX comprend des chaînes de mots discriminants qui s'adaptent à plusieurs descriptions entrées. En conséquence, le SLEX n'est pas aussi précis que le PLEX et ne sert que si le PLEX ne réussit pas à coder l'information.

En 1980, le volume du PLEX est passé d'environ 4 000 enregistrements à plus de 11 000 pendant la durée du recensement suédois. Le volume du SLEX a été maintenu à approximativement 500. Le degré de codage pour tout le recensement était de 71,5%, à peu près 68% pour le PLEX et 3% pour le SLEX. Le codage était effectué par comté et le degré de codage variait de 67% à 77% d'un comté à l'autre. Comme dans le cas de l'EDM, le coût de correspondance est négligeable. Dans l'exemple suivant, les descriptions relatives à un comté regroupant 314 529 personnes économiquement actives ont été opposées à un PLEX contenant 10 291 descriptions et à un SLEX comprenant 513 chaînes de mots. Le degré de codage total était de 74,7% et le coût s'élevait à environ 50 $.

Le système de codage du recensement suédois de 1985 était le même que celui utilisé en 1980. La seule différence notable : le volume du PLEX avait été augmenté à 20 000 et celui du SLEX à 1 900. Le degré de codage pour l'ensemble du recensement était de 73,7%, soit environ 69% pour le PLEX et 5% pour le SLEX. Le degré de codage des comtés variait de 72% à 75,8%.

Les remarques suivantes s'appliquent au codage automatisé de la profession dans les recensements. (1) Le codage automatisé coûte à peu près 10% de moins que le codage manuel. (2) Le codage se fait plus rapidement. (3) En raison de l'automatisation, la charge de travail du codage manuel est réduite et il est possible de n'employer que le personnel régulier au lieu d'engager temporairement un grand nombre de codeurs. Depuis quelques années, on utilise également le codage automatisé de la profession dans l'enquête sur le travail en Suède. Pour cette enquête, nous utilisons un dictionnaire contenant un peu moins de 5 000 descriptions. Le degré de codage est d'environ 70% avec de très faibles écarts mensuels.

On utilise également le codage automatisé pour d'autres enquêtes. L'exemple le plus remarquable est le codage des prêts de livres dans une enquête parrainée par le Swedish Authors' Fund. On a utilisé le codage automatisé pendant dix ans et au cours de cette période, le volume du dictionnaire est passé de 7 000 auteurs et titres de livres à plus de 65 000. Le degré de codage actuel est d'approximativement 75% et le système est très rentable comparativement à l'ancien système manuel.

APPLICATIONS DE NOUVELLES TECHNIQUES POUR LA PLANIFICATION RÉGIONALE ET LOCALE

Les statistiques ayant trait aux zones pour l'expansion urbaine, les zones récréatives, l'agriculture et l'utilisation des terres présentent beaucoup d'intérêt en Suède. L'information recueillie sert à fournir des statistiques utilisées pour la planification urbaine et régionale et pour la gestion des ressources naturelles. Ce genre de collecte de données serait pratiquement impossible sans l'accès aux techniques modernes. Les cartes, les photos aériennes et, dans une certaine mesure, l'information transmise par satellite servent de sources de données pour les statistiques sur l'utilisation des terres et les coordonnées permettant d'améliorer les variables géographiques.

Pendant quelques années, Statistics Sweden a expérimenté la télédétection (photos aériennes et données transmises par satellite). Les principales conclusions indiquent que pour l'instant, la télédétection ne peut satisfaire à des exigences de précision raisonnables et est trop coûteuse. Les méthodes d'enquête traditionnelles telles que l'échantillonnage et la collecte de donnée au moyen de questionnaires envoyés par la poste sont plus efficaces. Cependant, des projets spécifiques de télédétection vaudraient la peine d'être explorés. Par exemple, il faut 400 personnes pour estimer la production agricole. Chaque année, ces personnes moissonnent des zones sélectionnées d'un mètre carré de culture dispersées dans environ 50 000 champs. Il s'agit d'une procédure coûteuse et il est important de déterminer si la télédétection permettrait d'obtenir de l'information auxiliaire. Le système suédois d'assurance-récoltes constitue un autre exemple. Il est bien évident que ce système ne peut reposer entièrement sur les données transmises par satellite, mais l'on pourrait utiliser les photographies aériennes à titre de complément au système. Par exemple, les photos aériennes ont permis de documenter l'ampleur des inondations. On y a également recours pour délimiter les terres arables endommagées dans les cas où les agriculteurs réclament davantage de dommages-intérêts que ce qui est prévu d'après le classement de la zone. La télédétection pourrait également servir à recueillir de l'information sommaire sur les caractéristiques des récoltes et ainsi permettre un suivi manuel plus dirigé et plus efficace.

Beaucoup de statistiques sur l'utilisation des terres reposent sur les cadastres et les données transmises par satellite ne peuvent rivaliser avec ces cadastres en raison des coûts. Cependant, les registres ne fournissent pas d'informations sur les zones vertes et les étendues d'eau dans les régions à forte densité de population. Ce type d'informations pourrait être recueilli par satellite tout comme l'information sur l'expansion des régions à forte densité de population et les changements dans les zones d'accès commun. (L'accès commun est un privilège accordé par la loi suédoise qui permet aux individus de passer sur des terres appartenant à des particuliers sauf dans le cas de champs cultivés et de terrains privés près des résidences.)

Une évaluation générale indique que la télédétection est plus utile pour rassembler des statistiques sur l'utilisation des terres que des statistiques liées aux cultures. La plupart des travaux en matière de télédection sont encore au stade initial. Le ciel de la Suède n'est clair qu'un jour sur cinq en moyenne, et les systèmes conventionnels d'enquête utilisés à Statistics Sweden sont suffisamment efficaces pour justifier la lenteur du progrès dans l'utilisation de la télédétection (voir Hall-Knyves, 1988, et Justusson, 1984).

Règle générale, la nouvelle technologie est essentielle à la production de statistiques pour la planification régionale et locale (voir Ansen et al., 1988). Par exemple, les statistiques sur l'utilisation des terres dans les régions urbaines et leurs environs immédiats ont été élaborées grâce au traçage à l'aide de cartes à partir de photographies aériennes (voir Matérn, 1986). Dans les régions urbaines échantillonnées, l'utilisation des terres est déterminée au moyen du traçage sélectif à partir de photos aériennes. Les données des cartes existantes sur les régions urbaines et des cartes sur l'économie complètent l'information provenant des photos. Le traçage à l'aide de cartes se fait sur des tables lumineuses à partir des épreuves de photos aériennes en noir et blanc. On place un film de plastique transparent quadrillé (en centimètres) sur la photo et on trace les anciennes lignes de démarcation de la région urbaine sur le film. +A l'extérieur de ces lignes de démarcation, on trace les nouvelles limites extérieures de la zone d'accessibilité (distances réelles de 500 ou 1 000 mètres, d'après la taille de la population). Puis, on trace des points correspondant à des distances réelles de 200 mètres et moins. Le type d'utilisation des terres est codé directement sur le film de plastique selon un système spécial de classification. Il s'agit là d'un exemple d'enquête cartographique.

L'utilisation des coordonnées constitue un autre exemple de la nouvelle technologie. Tous les cadastres traditionnels sont actuellement remplacés par un système informatisé, processus qui devrait être achevé d'ici quelques années. Lorsque l'on établit de nouveaux cadastres, on y ajoute des coordonnées pour chaque unité de biens immeubles. Puis, on peut localiser les données liées aux biens immeubles pour les régions arbitrairement délimitées. L'analyse spatiale informatisée et la cartographie ouvrent de nouvelles avenues. +A l'avenir, il sera possible d'utiliser plus fréquemment les caractéristiques géographiques, ce qui s'avère important, la densité, les possibilités, la covariation spatiale et les autres variables géographiques étant si utiles pour les statisticiens et les planificateurs. Diverses agences suédoises collaborent à la production de cartes statistiques (de points, quadrillées et isarithmiques) et d'échelles informatisées.

QUESTIONS AYANT RAPPORT AU CARACTÈRE PRIVÉ ET À LA CONFIDENTIALITÉ

Au cours des vingt dernières années, Statistics Sweden a travaillé considérablement à préserver le caractère privé et la confidentialité des données recueillies et mémorisées. En 1973, la Suède faisait office de pionnier dans le domaine de la protection des données en adoptant une loi nationale sur les données assurant le caractère privé et la confidentialité dans les secteurs publics et privés. C'est principalement le débat public très animé sur les procédures de collecte de données et le contenu du recensement de la population et des logements de 1970 qui a conduit à l'adoption de cette loi. En effet, lors du recensement de 1970, on avait appliqué des méthodes désuètes de collecte de données où, par exemple, les concierges recueillaient les formules de recensement dans les immeubles de résidence. On avait également introduit une nouvelle section sur la scolarité. Dans certains pays, le revenu constitue une variable d'enquête

très délicate; en Suède, c'est l'instruction. Le recensement de la population et des logements de 1965 s'était très bien déroulé, l'agence n'ayant reçu que quelques appels téléphoniques pendant la période de collecte de données. En 1970, le standard a été bloqué pendant trois semaines au cours desquelles le personnel du bureau central du recensement n'a fait que rassurer des répondants inquiets et en colère. Bien que la grande majorité des ménages aient rempli leurs formules sans rappel ni suivi, Statistics Sweden a considéré qu'il fallait prendre des mesures immédiates pour remédier à ces nouveaux problèmes.

Statistics Sweden a donc commencé à élaborer un certain nombre de mesures de protection, seul ou avec d'autres agences et autorités. De plus, un nouvel organisme, le Data Inspection Board, a été formé afin d'exercer une surveillance sur les opérations de traitement des données des agences et organisations. Statistics Sweden n'a pas réduit ses activités relatives à la protection des données avec les années. Le public avait fait part de ses inquiétudes face au recensement de 1980. En 1983, une commission nommée par Statistics Sweden proposait de baser le recensement de 1985 sur des données provenant de dossiers existants plutôt que sur les données recueillies directement des ménages. Les médias, diverses organisations et un certain nombre de citoyens ont exprimé leurs inquiétudes face à cette proposition et le gouvernement a finalement décidé de ne pas modifier la méthode de collecte de données pour le recensement de 1985. La collecte de données du recensement de 1985 était à peine achevée qu'elle soulevait déjà de nouvelles protestations. En effet, en vertu de la loi sur le recensement de la population, le National Tax Board avait accès aux données relatives aux adresses afin de les comparer à celles qui figuraient dans ses propres registres. Cependant, avant que les ménages remplissent les formules de recensement, on les avisés que les données de recensement ne serviraient qu'à des fins statistiques. Comme la vérification des adresses par le National Tax Board faisait partie d'un plan de grande envergure pour poursuivre les fraudeurs de l'impôt, cette autorisation a été violemment critiquée.

En 1986, c'est au tour d'un projet suédois de recherche sociologique, une étude longitudinale nommée "Metropolitan", de déclencher un débat passionné. Ce projet était dirigé par un groupe de recherche de l'université de Stockholm avec une participation très minime de Statistics Sweden. Comme la plupart des données recueillies provenaient de dossiers administratifs et que bon nombre des 15 000 personnes n'étaient pas au courant de leur participation, la question de la confidentialité a été soulevée. Ce débat passionné, couvert quotidiennement par les médias, a duré six mois et, à l'instar des débats précédents sur les questions de la confidentialité, il a diminué l'empressement du public à participer aux enquêtes de Statistics Sweden. Le taux de non-répondants de l'enquête suédoise sur la population active a presque doublé pendant les premiers mois du débat. Statistics Sweden a dû déployer des efforts considérables pour arriver à réduire très modérément le taux de non-répondants et jusqu'à ce jour, certaines de ces enquêtes n'ont pu récupérer ces absences de réponses.

La révolution informatique a présenté certaines menaces pour la vie privée des particuliers en ce sens que nos capacités de recueillir, de copier, de stocker, de fusionner et d'extraire l'information sur les individus se sont accrues de façon considérable. Par ailleurs, c'est cette même technologie qui nous permet de protéger l'individu contre une surveillance des données potentiellement abusive. Dalenius (1988) fait une analyse des mesures de protection de l'intégrité des données d'enquête. Ces mesures ne sont pas seulement techniques, comme c'est le cas pour l'élimination de l'identité, la suppression, l'arrondi, etc., mais elles sont également légales et administratives.

Aujourd'hui, Statistics Sweden applique de façon courante un certain nombre de mesures protectrices. En vertu de sa politique, toute la matière ayant une valeur limitée pour les chercheurs ou pour laquelle on ne prévoit aucun besoin de rapprochement doit être identifiée. L'élimination de l'identité est considérée comme la garantie la plus sûre et la plus pratique. On peut l'appliquer dans le cadre de la collecte de données, lorsque le traitement des données est terminé ou après une période préalablement déterminée. La politique veut que les registres créés par rapprochement des dossiers ne soient pas identifiés et soient parfois même détruits après avoir été utilisés. La matière sur le choix des partis politiques, par exemple, est soumise à l'élimination de l'identité, malgré les restrictions que cela impose sur les recherches futures. On le fait pour des raisons de confidentialité uniquement et cet aspect de la politique devra être appliqué encore plus fréquemment à l'avenir. Nous tenons à souligner que l'élimination de l'identité n'est pas complètement sûre. Par ailleurs, l'identification amont est difficile à exécuter étant donné qu'elle nécessite des ressources considérables et de l'information détaillée sur les individus.

D'autres mesures sont de nature administrative; par exemple, on peut séparer le nom et l'adresse de l'individu de son numéro d'identification. Des mesures de sécurité limitent également l'accès à l'information et tous les visiteurs doivent s'identifier.

Un certain nombre de lois et de règlements sont également en application. Le Data Inspection Board doit donner son autorisation ou dans les cas simples délivrer un permis général avant que de nouveaux registres puissent être créés. Chaque employé doit signer une déclaration de secret professionnel et la violation de ce contrat entraîne des sanctions très sévères. L'information protégée ne peut être utilisée en cour par des témoins à moins que la nature de l'infraction soit telle que le tribunal décide de passer outre à ce règlement. Jusqu'à maintenant, les tribunaux de Suède n'ont jamais exercé ce droit. Statistics Sweden possède des systèmes d'accès qui limitent le nombre de personnes autorisées à travailler avec les différents registres. Avec le tout nouveau système d'accès Top Secret, seules certaines personnes ont accès aux registres et seulement à des terminaux préalablement déterminés. En théorie, on peut contourner le Top Secret en lisant les données directement à partir des disques ou en introduisant des erreurs au niveau des dispositifs de communication entre ordinateur et terminal. Toutefois, il s'agit là d'une méthode plutôt maladroite qui entraînerait des séries de nombres pratiquement incompréhensibles. Pour l'instant, nous considérons que le risque que représente l'introduction volontaire d'erreurs n'est pas assez important pour justifier des mesures de protection.

L'élimination totale de l'identité des éléments rend impossible les recherches futures telles que les études longitudinales et le rapprochement d'enregistrements. Nous avons besoin d'une technique qui rendrait possible l'élimination de l'identité tout en permettant le rapprochement d'enregistrements. Le chiffrement unilatéral des numéros d'identification personnels le permet. Une fois tous les numéros d'identification chiffrés, on peut traiter les données de nature confidentielle sans risquer qu'un individu soit associé aux données qui le concernent. Avec le chiffrement unilatéral, le numéro chiffré remplace le numéro d'identification et ne donne pratiquement aucune information. Cela suppose qu'il n'existe aucun autre moyen d'identification dans le registre.

La première étape du chiffrement des numéros d'identification personnels consiste à établir une correspondance avec un ensemble de numéros chiffrés à l'aide d'un algorithme complexe. Cet algorithme est inversement non équivoque, ce qui signifie que deux numéros d'identification différents correspondront toujours à deux numéros chiffrés différents.

L'algorithme devrait être tel qu'il serait pratiquement impossible de décoder le numéro chiffré. Conformément à la loi sur la protection des données, le registre chiffré n'est plus un registre d'individus. Néanmoins, il demeure toujours confidentiel en raison du risque minime que représente l'identification amont ou toute autre identification techniquement difficile.

Après le chiffrement, le dossier non chiffré est classé aux archives nationales et il n'en reste aucun vestige à Statistics Sweden. Le support d'information qui contenait le dossier non chiffré est même démagnétisé.

On peut créer un algorithme de chiffrement unilatéral de plusieurs façons. Statistics Sweden a choisi d'utiliser un algorithme conventionnel capable de coder et de décoder à la même vitesse. L'algorithme et son inverse sont ultra-secrets. Il faut franchir plusieurs étapes avant d'arriver au programme de chiffrement définitif, qui est un programme de chargement protégé par le nouveau système d'autorisation. Les détails sur la méthode de chiffrement unilatéral sont exposés dans Rapaport et Block (1986). Récemment, le Data Inspection Board a autorisé Statistics Sweden à expérimenter le chiffrement unilatéral dans deux enquêtes, sur une période de trois ans.

Statistics Sweden a eu un programme de recherche sur les statistiques et la confidentialité pendant plusieurs années et le bureau a collaboré avec beaucoup d'autres agences et organisations en Suède et à l'étranger. Deux conférences internationales sur le caractère privé et la confidentialité de l'information ont été tenues, en 1978 et en 1987. Statistics Sweden (1988) réunit de la documentation sur la dernière conférence.

DIFFUSION

Statistics Sweden applique deux grands principes à la diffusion de ses statistiques. Le premier c'est que les statistiques devraient être facilement accessibles à l'utilisateur et être adaptées à sa capacité de les utiliser. Le deuxième c'est que le citoyen moyen devrait être tenu au courant des développements importants dans notre société. Statistics Sweden atteint le grand public en envoyant des exemplaires de ses publications dans les bibliothèques et en publiant des communiqués de presse qui sont diffusés par les médias. Nous n'offrons aucune rémunération pour la diffusion des statistiques de l'agence.

Au contraire, nous exigeons souvent que les utilisateurs paient les publications et les statistiques présentées sur d'autres supports. Cela permet de limiter quelque peu la demande pour ces statistiques et aide à payer les frais de diffusion. Notre politique ne consiste pas à demander le prix le plus élevé possible, mais bien à atteindre le niveau optimal de diffusion. La diffusion électronique coûte cher étant donné qu'il faut couvrir les frais de recherche et de développement et que les mêmes statistiques sont souvent disponibles sous des formes moins coûteuses, les publications, par exemple. Un certain degré de diffusion se fait gratuitement, notamment dans les municipalités et autres organismes et agences gouvernementales. Ceux qui participent aux enquêtes reçoivent habituellement un exemplaire de la publication qui en résulte afin qu'ils sachent à quoi ils ont participé.

Publications

Statistics Sweden produit annuellement 700 publications régulières totalisant 600 000 numéros. Un nombre sans cesse croissant de ces publications sont réalisées au moyen de techniques de traitement de texte évoluées. La politique officielle veut que la production de statistiques soit aussi décentralisée que possible et l'étape de la publication n'échappe pas à cette règle. Nous visons également à remplacer un certain nombre de publications par des supports électroniques.

Supports électroniques

Statistics Sweden utilise les supports électroniques suivants pour mettre les statistiques à la disposition du public : bandes magnétiques, disquettes, disques compacts - ROM, vidéotex et bases de données interactives. Chaque support possède des avantages et des inconvénients et peut compléter ou remplacer les rapports publiés.

Bandes magnétiques

Moyennant une somme modique, Statistics Sweden publie des données sur bandes magnétiques. Certains utilisateurs reçoivent des bandes chaque mois. Toutefois, la demande n'est pas très forte pour les données sur bandes magnétiques et ce support ne peut être considéré comme un mode de diffusion important.

Disquettes

Au cours des dernières années, nous avons diffusé des données sur disquettes, mais seulement à petite échelle. Jusqu'à maintenant, il n'existe aucun principe ou norme pour rendre les données disponibles sur disquettes. Il n'existe pas non plus de logiciel adapté au traitement des disquettes de Statistics Sweden.

Disques compacts - ROM (à mémoire morte)

La diffusion des données via les disques compacts - ROM est encore au stade expérimental. On vise à stocker d'importants ensembles de données sur disques compacts. + A l'heure actuelle, quelques bibliothèques suédoises peuvent prendre les disques compacts en charge et le système scolaire suédois s'est montré intéressé aux importants fichiers stockés sur disques compacts aux fins de diffusion dans les écoles.

Vidéotex

On peut décrire le vidéotex comme un système d'information statistique, c'est-à dire des tables et des diagrammes accompagnés de notes explicatives et d'analyses constituées de façon telle que l'utilisateur peut parcourir le système à l'aide de menus pour extraire l'information qui l'intéresse. Ce système est purement visuel bien que le terme "vidéo" puisse amener à penser qu'il est également sonore. Au cours des trois dernières années, nous avons éprouvé le

vidéotex à grande échelle. Les écoles suédoises ont manifesté beaucoup d'intérêt et un système est présentement utilisé, mais à une échelle réduite par rapport au test. Le système actuel contient environ 50 pages différentes de vidéotex.

Bases de données

Depuis le début des années 80, Statistics Sweden utilise les bases de données statistiques comme autre solution aux publications. Parmi les options de diffusion électronique offertes, les plus utilisées sont les bases de données, la base de données à séries chronologiques et la base de données statistiques régionale. Ces bases de données emploient le gestionnaire de base de données AXIS et l'utilisateur communique en direct avec l'ordinateur de Statistics Sweden pour générer des tables adaptées à ses besoins. Nous tenons à souligner que AXIS est plus souple que tout autre gestionnaire de base de données. Les données sont stockées de façon à permettre la génération de tables conçues par l'utilisateur. Cependant, le système AXIS comporte un inconvénient : il ne peut jamais être utilisé comme base de données pour du texte intégral. Comme l'utilisateur crée les tables, celles-ci ne sont pas stockées dans la base de données et par conséquent, aucun commentaire important, à l'exception des notes en bas de page expliquant les variables, etc., ne peut être mémorisé. Cette caractéristique empêche AXIS de remplacer complètement la publication des séries de rapports réguliers de Statistics Sweden, intitulés Statistical Reports, étant donné que seule une base de données à texte intégral peut comprendre les commentaires et l'analyse essentiels à un rapport statistique.

Le système de base de données actuel de Statistics Sweden a fait son apparition en 1982. Après trois ans, on comptait 110 abonnés en direct. Six ans plus tard, ce chiffre a doublé pour atteindre 220. On estime à environ 500 le nombre d'utilisateurs réels. Le nombre de matrices stockées est évalué à 1 200, ce qui correspond à près de 150 mètres de documentation publiée en librairie. Les bases de données gérées par AXIS contiennent approximativement 30% des statistiques publiées dans les rapports. Certains sujets ne sont pas du tout représentés dans les bases de données.

Le temps machine que les abonnés à la base de données achètent de Statistics Sweden pour traiter les données diminue radicalement. Bon nombre d'abonnés possèdent leurs propres OP qui leur servent à extraire les données voulues de la base de données, à les traiter et à les analyser.

L'exploitation des bases de données coûte environ 10 millions de couronnes par année (2 millions de dollars canadiens). Le revenu provenant des utilisateurs s'élève à environ 1,5 million par année. Néanmoins, nous nous sommes fixé comme objectif de couvrir les coûts d'exploitation grâce aux revenus générés par les bases de données. Comme les bases de données constituent un nouveau support, on s'attend à ce qu'à la longue, elles deviennent financièrement indépendantes. Toutefois, cette exigence ne s'applique pas aux publications. Afin de réduire l'écart entre le revenu et le coût des bases de données, on peut proposer d'augmenter de façon considérable les frais d'abonnement. La société a de plus en plus recours à l'électronique et une plus grande partie de ce qui est maintenant publié sous forme de rapports sera transférée dans les bases de données; à long terme, celles-ci bénéficieront du soutien financier actuellement accordé aux publications.

Bases de données à texte intégral

Nous n'avons pas encore utilisé les bases de données à texte intégral comme support de diffusion. Cependant, il y a énormément de matière adaptée à la diffusion au moyen des bases de données à texte intégral. Les statistiques, dont la valeur dépend de la rapidité de diffusion, constituent la matière la mieux adaptée à ce type de support. Les communiqués de presse et les rapports statistiques, du moins certains d'entre eux, en sont des exemples. Chaque année, nous publions approximativement 400 communiqués de presse totalisant 220 000 parutions. On réaliserait des économies énormes si les communiqués de presse étaient diffusés par l'entremise de bases de données à texte intégral interactives.

Une mise en garde s'impose au sujet de la diffusion par l'entremise des bases de données statistiques et d'autres supports électroniques. Les publications statistiques renferment beaucoup d'informations qui devraient être comprises sur les supports électroniques, notamment un type quelconque d'instructions relatives à la qualité des données, le nom d'une personne ressource qui peut répondre aux questions et même, dans certains cas, de l'informatique technique sur la méthode de collecte et de traitement des données. Il y a toujours le risque que les chiffres présentés dans une table ou un diagramme attrayants soient perçus comme étant incontestables. Il y a aussi une opinion fausse selon laquelle l'information non quantitative ne constitue pas vraiment de l'information. Nous, qui travaillons dans les agences nationales, savons qu'il y a des données sûres et d'autres qui ne le sont pas. Mais comment le public exposé à des sommes croissantes de données peut-il juger de leur qualité? L'élaboration des supports électroniques et leur disponibilité ne cessant de croître, il faut résoudre le problème des instructions sur la qualité des données.

DÉVELOPPEMENTS FUTURS

On reconnaîtra deux importants développements au niveau du travail d'enquête à Statistics Sweden. Premièrement, les ordinateurs ont pris en charge la totalité ou une partie des opérations. Dans le présent document, nous avons examiné l'automatisation du codage et de l'interview. Le degré varie de l'automatisation quasi complète aux diverses formes d'assistance par ordinateur. Deuxièmement, on tente de fusionner plusieurs opérations afin de réduire leur nombre. D'un point de vue, ces développements ne présentent aucun problème. L'agence peut réduire ses coûts et son personnel tout en accomplissant davantage de tâches, y compris des tâches nouvelles. Toutefois, d'un autre point de vue, le progrès technologique crée des problèmes. Les systèmes automatisés doivent préserver la confidentialité, être simples, bien éprouvés et extrêmement fiables. Autrement dit, l'automatisation doit être à toute épreuve. L'intégration signifie qu'une personne exécute souvent les tâches auparavant assumées par plusieurs. Et même si la nouvelle technologie peut simplifier certaines tâches, elle exige de nouvelles compétences de la part du personnel.

Les récents développements et tendances vers la décentralisation ont conduit à ce que l'on pourrait appeler la révolution de l'OP. Il faut donc une méthode entièrement nouvelle d'éducation et de formation pour notre personnel. A cette fin, Statistics Sweden a lancé un vaste programme d'éducation permanente pour ses employés. Le programme comprend des cours dans différents domaines et la formation est également offerte aux gens de l'extérieur de l'agence. Une série de cours s'adresse aux utilisateurs de statistiques tandis qu'une autre est destinée aux producteurs. A notre connaissance, aucun autre bureau de statistiques au monde

n'offre un équivalent à ce programme, qui varie des mini-cours de deux heures au programme complet de Ph.D. avec ses propres enseignants dont deux titulaires de chaire.

Les préoccupations relatives à la qualité et aux coûts de même que les facteurs humains doivent absolument dominer la technologie. Pendant près de dix ans, Statistics Sweden a subi d'importantes restrictions budgétaires et la tendance a été de remplacer le personnel par la technologie. Pendant cette période d'adaptation, on a mis l'accent sur la production de quantités égales ou mêmes supérieures de statistiques avec un personnel réduit et des fonds plutôt inadéquats. On a effectué très peu de travail au niveau des études méthodologiques telles que l'évaluation, les préenquêtes et les expériences, sans oublier le contrôle de qualité continu. On en a fait davantage au début des années 1970. Toutefois, il semble que les priorités vont changer de nouveau. Tous les programmes statistiques font actuellement l'objet d'un examen minutieux qui mènera à des remaniements et à des procédures améliorées. Le personnel de Statistics Sweden s'est dans une large mesure adapté à l'utilisation répandue des OP, ce qui devrait lui permettre de se consacrer davantage aux questions relatives à la qualité.

HYPERQUEST: UNE NOUVELLE MÉTHODE POUR INTERVIEWER À L'AIDE DES ORDINATEURS PORTATIFS

M. A. JARO,

Systems Automation Corporation

RÉSUMÉ

Cet article décrit un nouveau système qui combine le processus d'interview avec une base de données hyper-relationnelle. Le module d'interview offre plusieurs possibilités (choix multiples tirés de listes fixes ou variables, possibilité de réponses longues, retour à n'importe laquelle des questions précédentes, fichiers d'assistance, édition, boucles, questions à trajet complexe, etc.). Au lieu d'être placées dans un fichier séquentiel, les réponses le sont dans une base de données hyper-relationnelle. La nature de cette base de données est telle que toutes les données sont automatiquement reliées jusqu'au niveau des zones. Un processeur spécial d'interrogation peut être utilisé pour parcourir l'information de façon structurée ou encore pour effectuer des consultations ponctuelles. Ce processeur est un système orienté vers l'objet qui permet à tout usager (même s'il ne possède aucune connaissance préalable du contenu ou de la structure de la base de données) d'effectuer des interrogations complexes et de générer des rapports. Il est également possible d'obtenir des informations statistiques sur les réponses. Ce concept est très différent de celui utilisé avec les systèmes de bases de données standards, car dans ceux-ci les interrogations sont lentes, exigent un apprentissage difficile (langage d'interrogation structuré) ou encore sont impossibles si l'on ne connaît pas de façon détaillée les relations existant entre les entités de la base de données.

1. INTRODUCTION

La collecte de données assistée par ordinateur (CADAC) et l'interview personnelle assistée par ordinateur (CAPI) prennent de plus en plus d'importance par suite de l'arrivée des ordinateurs portatifs supercompacts. La technologie actuelle du matériel permet maintenant à un intervieweur de venir faire l'interview avec un ordinateur léger alimenté par piles. Les facteurs suivants ont contribué à faire de la CAPI une technique réalisable et attrayante sur le plan de la technologie : amélioration du contraste et de la résolution des écrans, disques rigides peu coûteux, augmentation de la capacité des disquettes, mémoires internes de grande taille, alimentation par piles et prix relativement faible.

Les avantages de la CAPI sur le plan technique sont évidents : édition en ligne assurant que les données sont complètes et cohérentes, ce qui réduit les risques d'avoir à faire d'autres visites par suite d'échecs au contrôle comme cela se produit avec les gros ordinateurs; comme l'ordinateur pose les questions pertinentes, il est possible d'introduire dans le questionnaire des branchements complexes ainsi que plusieurs trajets sans compliquer la formation; les coûts d'impression et les problèmes de conception associés aux questionnaires sur papier peuvent être éliminés; dans le cas des enquêtes donnant lieu à des visites multiples, les réponses obtenues lors des visites précédentes peuvent être emmagasinées et la cohérence longitudinale peut être vérifiée grâce au processus d'édition; les coûts de saisie des données ou de lecture optique sont réduits car les résultats peuvent être lus par la machine; on peut faire une utilisation plus libérale des réponses écrites, car ce sont les intervieweurs eux-mêmes qui entrent les données, plutôt que des préposés à la saisie essayant d'interpréter ce que les intervieweurs ont écrit.

Il est regrettable que les logiciels n'aient pas progressé au même rythme que le matériel. Les systèmes actuels d'interview sont assez difficiles à utiliser et créent en sortie des fichiers séquentiels. Les images-écrans sont plutôt confuses, les opérations sont complexes et la consultation des résultats est généralement impossible ou au mieux inflexible.

Pour pallier à ces difficultés, nous avons développé le système HyperQuest. Il est composé d'un module d'interview, d'une base de données hyper-relationnelles et d'un système d'édition, qui sont tous trois incorporés dans un même logiciel pouvant être chargé sur un

ordinateur portatif. Cet article décrit certains des aspects de ce système qui le rendent tout à fait unique en son genre.

2. LE MODULE D'INTERVIEW

2.1 L'interface de l'intervieweur

Une des exigences principales concernant les systèmes d'interview est la facilité d'utilisation. Les images-écrans encombrées de questions multiples et d'informations non-pertinentes sont difficiles à utiliser par les intervieweurs. Il ne faut jamais oublier que les intervieweurs ne possèdent en général aucune expérience informatique et sont souvent intimidés par l'idée d'employer des ordinateurs. Pour cette raison, nous avons choisi un format d'images-écran composé seulement d'un cadre pour les questions, d'un cadre pour les réponses et d'une ligne d'état. Ce format reste le même quelle que soit la question et constitue donc une interface simple et dégagée, présentée toujours de la même façon à l'intervieweur.

Les réponses à choix multiples sont sélectionnées à l'aide d'un curseur mobile. Des images-écrans d'assistance peuvent être créées pour chaque question, afin que l'intervieweur qui ne se rappelle pas des règles associées à une réponse puisse obtenir une ou plusieurs images-écrans d'assistance en pressant la touche de fonction F1. Les images-écrans d'assistance peuvent être facilement élaborées à l'aide d'un éditeur spécial.

L'intervieweur peut interrompre l'interview en pressant une seule touche de fonction. Les résultats sont alors automatiquement sauvegardés dans le système. Si l'intervieweur reprend l'interview, le système recommence automatiquement l'exécution au point où elle avait été abandonnée.

Dans les interviews complexes, il est souvent nécessaire que des cartes d'information soient montrées au répondant. Il est possible de créer ces cartes à l'écran avec l'éditeur des images-écrans d'assistance et de les afficher automatiquement lorsque cela est nécessaire.

2.2 Consultation des données entrées par l'intervieweur

Il est important que l'intervieweur soit capable de revenir à n'importe quelle question qu'il a déjà posée pour éventuellement en modifier la réponse. De telles modifications peuvent ensuite avoir un effet sur le trajet de l'interview. Le module d'interview HyperQuest permet automatiquement de revenir à la question précédente ou à n'importe quelle autre question (la liste est affichée à l'écran). Si une question se trouve dans une boucle, l'intervieweur peut revenir à n'importe quelle version de cette question. Lorsqu'il revient en arrière pour afficher une question déjà posée et sa réponse, il dispose de plusieurs options : 1) il peut repasser au point où il se trouvait auparavant; 2) il peut modifier la réponse et avancer automatiquement soit à la dernière question posée ou encore à la première question à poser si le trajet de l'interview a été modifié par la nouvelle réponse; 3) il peut faire défiler les questions de l'interview une à la fois.

M. William L. Nicholls et Mme Carol House ["Designing Questionnaires for Computer-Assisted Interviewing: A Focus on Program Correctness"; U.S. Bureau of the Census, Third Annual Research Conference, Proceedings, 1987] ont traité des problèmes qui se rapportent au déplacement vers l'arrière et au déplacement vers l'avant qui en résulte. Nous avons utilisé les directives qu'ils proposent dans la conception du module d'interview HyperQuest. En outre, cet article présente divers principes qu'il est bon de suivre pour concevoir un bon outil d'enquête.

En pratique, le module de l'intervieweur conserve une pile interne renfermant les questions posées. Les déplacements vers l'arrière font revenir la base de données au moment de la question affichée et toutes les questions qui lui ont fait suite sont placées dans une pile en suspens. S'il y a lieu, des réponses seront tirées de cette pile en suspens. Grâce à cette fonction, on évite de poser à nouveau des questions dont on a déjà obtenu les réponses.

2.3 Le langage de programmation Script

Le processus d'interview est piloté par un langage puissant de la quatrième génération appelé SPL (Script Programming Language). Ce langage possède diverses caractéristiques spéciales permettant le codage d'interviews très complexes. Voici une liste de certaines de ces caractéristiques :

Des variables de liste spéciales et des sous-programmes de traitement de liste permettent de traiter les choix des menus, et les interrogations pouvant donner lieu à plusieurs réponses (questions à choix multiples).

Des fonctions d'assistance programmables sont disponibles pour aider l'usager final à effectuer une interview.

Les interviews peuvent être interrompues n'importe quand et redémarrées automatiquement à la question suivante.

Les branchements aux questions peuvent être déterminés à partir du contenu de la base de données, ce qui permet de poser des questions intelligentes, de vérifier l'uniformité des réponses et de composer dynamiquement des questions.

Des fonctions d'édition puissantes permettent aux usagers d'éditer les entrées longues. En outre, diverses fonctions permettent de manipuler les réponses de l'usager.

Les options des menus peuvent être tirées de fichiers ASCII ordinaires, ce qui permet de modifier facilement les séquences types.

Des fichiers peuvent être écrits pour que l'on puisse réaliser des interviews et faire la collecte des données à base d'enregistrements.

L'intervieweur peut revenir à n'importe quelle question précédente et reprendre le déplacement vers l'avant sans avoir à entrer à nouveau les réponses, et cela quelle que soit la complexité du trajet à l'intérieur du questionnaire.

2.4 Le langage auteur

Un langage auteur spécial et un processeur ont été rédigés afin de simplifier grandement l'élaboration des interviews. Le langage auteur sert à interviewer le concepteur du questionnaire au sujet de l'interview qu'il désire concevoir. Le code SPL correspondant à l'interview cible est ensuite automatiquement généré. Cela simplifie l'élaboration des interviews et contribue à assurer l'exactitude de l'outil d'enquête.

La plupart des types de questions sont disponibles par l'intermédiaire du langage auteur. Ce sont entre autres les réponses numériques, les réponses courtes (un nom par exemple), les réponses longues (maximum de 20K de caractères avec édition de type traitement de texte), et les réponses à choix multiples suivantes : 1) liste fixe; 2) liste fixe avec réponse "autre"; 3) liste tirée de la base de données; 4) liste à laquelle l'usager peut ajouter des éléments; 5) liste avec plusieurs réponses. Les listes de choix multiples peuvent être emmagasinées soit

dans la base de données ou dans un fichier simple ASCII. Le langage auteur permet l'utilisation de branchements conditionnels et de boucles. Les boucles sont utilisées dans les sous-sections du questionnaire (p. ex., pour chaque enfant, répondez à la question suivante : ---). Les questionnaires sur papier et la plupart des systèmes d'interview limitent de telles boucles (p. ex. maximum de sept enfants). HyperQuest n'impose aucune limite de ce genre à l'usager et n'utilise pas d'espace sur le disque pour les questions sans réponse ou pour les zones laissées en blanc lorsque le nombre d'entrées est inférieur au maximum admissible.

3. LA BASE DE DONNÉES HYPER-RELATIONNELLE

Cette section décrit le concept des bases de données hyper-relationnelles et indique comment il diffère de celui des bases de données classiques. Les exemples fournis ne se rapportent à aucun type d'enquête particulier et n'exigent donc aucune connaissance des techniques d'enquête pour être compris.

3.1 Le concept de la base de données hyper-relationnelles

HyperQuest utilise un modèle de base de données complètement nouveau : le modèle hyper-relationnel. Ce modèle présente les avantages suivants par rapport à celui des bases de données relationnelles classiques :

Les relations sont définies au niveau des éléments d'information (zones). Cela élimine les problèmes de normalisation de données propres à la mise à jour des tables de données.

Des zones et des relations peuvent être ajoutées ou supprimées sans qu'il soit nécessaire de réorganiser la base de données.

Des opérations complexes telles que les liaisons peuvent être effectuées à l'aide d'un système d'interrogation orienté vers l'objet qui n'exige pas que l'usager connaisse un langage d'interrogation tel que SQL.

Toutes les relations entre les entités sont prises en charge automatiquement. Ce sont, entre autres, les relations bi-univoques, co-univoques, univoques et multivoques ainsi que la répétition des sous-groupes.

Le repositionnement et la réindexation des données sont simplifiés car la base de données ne renferme que quelques fichiers physiques.

La même base de données peut renfermer des applications diverses. Cela facilite l'établissement de liens entre des applications ou des sous-problèmes différents sans qu'il soit nécessaire de faire appel à des systèmes répartis complexes.

Aucun schéma prédéfini n'est nécessaire -- les relations entre les éléments d'information peuvent être définies à l'aide d'un dictionnaire simple.

Le problème des données manquantes est éliminé car aucune donnée ne manque dans la base de données.

Les interrogations sont extrêmement rapides.

3.2 Le modèle hyper-relationnel

3.2.1 La relation

Au lieu de faire appel à des enregistrements logiques comme dans la plupart des systèmes de base de données, le modèle hyper-relationnel utilise le concept de la relation. Une relation sert à emmagasiner et relier entre eux les éléments d'information. Une relation est composée d'un attribut et de deux arguments, et constitue donc un triplet. Examinons les exemples suivants de relations :

LANGUE (DANIEL SAINT-PIERRE, FRANÇAIS)

Cette relation indique que la LANGUE de DANIEL SAINT-PIERRE est le FRANÇAIS.

ÂGE (GEORGE, 42)

L'ÂGE de GEORGE est de 42 ans.

CHEF (MICHEL,DANIEL)

MICHEL est le CHEF de DANIEL.

Cette méthode de représentation de l'information a été utilisée par les systèmes d'intelligence artificielle et ces triplets sont dits OAV (objet, attribut, valeur); toutefois, ce concept a été considérablement développé dans notre modèle grâce à l'emploi de symboles et d'objets.

3.2.2 Symboles

Les arguments d'une relation peuvent être des valeurs numériques, des chaînes de texte ou des symboles. Un symbole est un argument qui est soit un mot ou un nom. Les symboles sont indexés et permettent d'explorer la base de données. Grâce à cette méthode d'indexation, il est facile de trouver toutes les personnes dont MICHEL est le chef ainsi que tous les chefs de DANIEL, l'ÂGE de toutes ces personnes, toutes les relations associées à GEORGE, etc.

3.2.3 Objets

Un objet est un ensemble de symboles représentant un type d'entité correspondant à une situation réelle. Par exemple, dans les relations suivantes :

LANGAGE (DANIEL SAINT-PIERRE, COBOL)
LANGAGE (DANIEL SAINT-PIERRE, ADA)
LANGAGE (MICHEL BÉRUBÉ, COBOL)
LANGAGE (MICHEL BÉRUBÉ, FORTRAN)

COBOL, FORTRAN et ADA sont tous des langages de programmation. Daniel Saint-Pierre et Michel Bérubé sont des noms d'employés. L'ensemble des langages de programmation peut constituer un objet ayant par exemple pour nom [LANGAGE-PROGRAMMATION].

De même, l'ensemble des employés peut constituer un objet ayant par exemple pour nom [EMPLOYÉ].

Un objet renferme un type d'élément. L'objet [LANGAGE-PROGRAMMATION] est semblable à une boîte renfermant une série de cartes. Le nom d'un langage de programmation est inscrit sur chacune de ces cartes. Il est possible d'ajouter des cartes à cette boîte ou d'en retirer.

3.2.4 Définitions relationnelles

Un dictionnaire de données définit les relations à l'intérieur d'une base de données en fonction des objets qu'elle renferme. Par exemple, une entrée de dictionnaire de la relation LANGAGE utilisée ci-dessus serait :

LANGAGE ([EMPLOYÉ], [LANGAGE-PROGRAMMATION])

Cette ligne spécifie que l'objet [EMPLOYÉ] et l'objet [LANGAGE-PROGRAMMATION] sont liés à l'aide de la relation LANGAGE.

3.2.5 Exploration de la base de données et interrogations

Examinons les deux types de définitions relationnelles suivants :

LANGAGE ([EMPLOYÉ], [LANGAGE-PROGRAMMATION])
LANGAGE-PROJET ([PROJET], [LANGAGE-PROGRAMMATION])

La relation LANGAGE définit les langages que les employés connaissent et la relation LANGAGE-PROJET définit les langages utilisés dans les projets. Pour identifier les employés connaissant les langages à utiliser dans certains projets, on doit exécuter les opérations suivantes :

Création d'un objet [EMPLOYÉ]
à partir d'un objet [LANGAGE-PROGRAMMATION]
à partir d'un objet [PROJET]
à partir d'une sélection de projets.

L'usager peut spécifier facilement cette interrogation en sélectionnant des choix dans un système piloté par menu. L'exécution se déroule dans le sens inverse :

l'usager sélectionne les projets qui l'intéressent;
les langages de programmation devant être utilisés dans les projets sélectionnés sont extraits (relation LANGAGE-PROJET);
les employés connaissant les langages de programmation obtenus sont extraits (relation LANGAGE).
On obtient ainsi une liste des employés ayant les attributs recherchés.

En général, on peut utiliser des objets pour en créer de nouveaux (établissement d'une liste des employés connaissant COBOL et ADA), pour limiter des objets (limitation de la liste des employés à ceux ayant travaillé à certains projets), pour ajouter des objets (ajout d'employés à la liste des employés ayant une maîtrise), pour éliminer des objets (élimination des employés ayant travaillé sur le projet x). Ces opérations sont équivalentes aux opérations logiques ET, OU et NON ET, mais on évite ainsi d'embrouiller l'usager avec une telle terminologie.

3.2.6 Création des relations

L'élaboration d'une interview au moyen d'un langage auteur ou directement à l'aide de SPL crée des relations pouvant être consultées. Par exemple, une question sur l'état matrimonial peut produire la relation :

ÉTAT MATRIMONIAL ([répondant], (état))

On peut avoir :

ÉTAT MATRIMONIAL (345, CÉLIBATAIRE)
ÉTAT MATRIMONIAL (345, DIVORCÉ)

Les questions à réponses multiples donnent lieu à des relations multiples. Par exemple, une question sur les saveurs préférées de crème glacée peut donner les réponses suivantes :

SAVEUR (345, VANILLE)
SAVEUR (345, CERISE)
SAVEUR (345, PÊCHE)
SAVEUR (89, CERISE)

Le répondant 345 aime les saveurs à la vanille, aux cerises et aux pêches tandis que le répondant 89 n'aime que la saveur aux cerises. L'entrée de dictionnaire de cette relation pourrait être :

SAVEUR ([répondant], [saveurs_préférées]).

À partir de cela, il est possible de déterminer quels sont les répondants qui aiment la crème glacée aux cerises ou encore les saveurs préférées du répondant 345.

4. LE MODULE DE SORTIE

En plus de la fonction d'interrogation, l'obtention de données statistiques et l'établissement de rapports constituent des éléments essentiels d'un bon système d'enquête.

Des données statistiques telles que la moyenne, l'écart-type, le minimum et le maximum peuvent être obtenues pour les données numériques à l'aide du processeur d'interrogation. On prévoit aussi créer une fonction similaire d'analyse des données catégorisées.

Un programme d'édition est incorporé au système d'interrogation. L'usager établit à l'écran la forme du rapport qu'il désire obtenir. Il peut sélectionner les zones et intégrer au rapport des en-têtes et des titres en bas de page. Des groupes répétitifs peuvent être définis et imprimés. Les rapports peuvent être sauvegardés et extraits en fonction des besoins. Il est possible de produire des rapports assez complexes facilement et très rapidement.

Parfois, il est nécessaire d'obtenir des enregistrements séquentiels afin de les entrer dans un autre programme. SPL peut être utilisé pour écrire les données (en tout ou en partie) dans un fichier séquentiel dans le format voulu.

5. CONCLUSION

Le fait d'avoir combiné un module d'interview conçu expressément pour les enquêtes à une base de données hyper-relationnelle dotée de fonctions d'interrogation et d'édition représente un progrès considérable en ce qui concerne la technologie de l'interview assistée

par ordinateur. Le but recherché était de fournir au concepteur d'interviews un outil puissant, et à l'intervieweur, un outil extrêmement simple. HyperQuest n'est pas un système statique. Nous pouvons y ajouter de nouvelles fonctions pour répondre aux besoins des applications, et toutes les suggestions des usagers actuels et potentiels sont toujours bien accueillies.

Matthew Jaro est directeur de la recherche et du développement
System Automation Corporation
8555 - 16th Street
Silver Spring, Maryland 20910
U.S.A.

(301) 565-9400

L'IMPACT DE LA TECHNOLOGIE DE POINTE SUR LA COLLECTE DES DONNÉES

W. L. NICHOLLS II,

US Bureau of the Census

L'IMPACT DE LA TECHNOLOGIE DE POINTE SUR LA COLLECTE DES DONNÉES

William L. Nicholls II
Field Division, U.S. Census Bureau[1]

La présente communication traite de l'impact de la technologie de pointe sur la collecte des données d'enquête. Bien qu'elle vise une période assez étendue, de 1970 à 2000, sa portée est limitée de deux façons.

Premièrement, cette communication porte surtout sur l'effet de la technologie de pointe sur la collecte des données effectuée par des organismes gouvernementaux. L'utilisation que font les organismes du secteur privé des mêmes technologies est mentionnée surtout comme une étape menant à leur adoption par les organismes gouvernementaux qui s'occupent de collecte des données.

Deuxièmement, la communication porte principalement sur les progrès réalisés aux États-Unis et plus particulièrement sur ceux qui se sont produits au U.S. Census Bureau. L'accent mis sur ces aspects n'a pas pour but de minimiser les travaux semblables effectués ailleurs, cela reflète simplement le fait que l'auteur connaît mieux les travaux effectués plus près de chez lui.

Dans cette communication, l'incidence de la technologie de pointe sur la collecte des données d'enquête est divisée en deux phases.

La phase 1 comprend les innovations technologiques introduites dans la collecte des données d'enquête avant 1980. On s'intéresse surtout à l'interview téléphonique assistée par ordinateur (ITAO), méthode utilisée tout d'abord dans le secteur privé au cours des années 70. On mentionne aussi brièvement deux méthodes associées, soit les sondages téléphoniques à composition aléatoire (STCA) et l'interview téléphonique centralisée.

Les technologies de la phase 2 comprennent les innovations dont les premières applications importantes en matière de collecte de données ont été faites après 1980. Ces technologies comprennent des sous-produits de la révolution micro-informatique, comme l'interview sur place assistée par ordinateur. Elles comprennent aussi de nouveaux progrès tels que la technologie de la parole et les appareils portatifs de reconnaissance de caractères manuscrits qui pourraient être utilisés de façon importante dans la collecte des données d'enquête d'ici l'an 2000.

Une évaluation des conséquences qui apparaissent à la suite de l'utilisation des technologies de la phase 1 pour la collecte des données d'enquête peut mieux nous préparer à faire face aux défis que présentent les technologies de la phase 2.

1. Technologies de la phase 1

L'interview téléphonique assistée par ordinateur (ITAO) constitue le premier exemple important de l'incidence de la technologie de pointe sur la collecte des données d'enquête. Cette méthode a permis d'appliquer l'informatique interactive aux tâches de collecte des données qui, depuis que des enquêtes sont menées par interview, étaient effectuées manuellement. Non seulement la méthode de l'ITAO combine la saisie des données avec la collecte des données, elle permet d'appliquer la puissance qu'offre l'informatique interactive

[1]/Les opinions exprimées dans cette communications sont celles de l'auteur et ne représentent pas nécessairement les opinions ou les politiques officielles du U.S. Census Bureau.

pour tenir les dossiers sur le terrain, pour faire l'ordonnancement des appels et des rappels, pour adapter les questions à la situation, pour poser au besoin des questions d'approfondissement, pour commander l'enchaînement des questions et pour relever et clarifier les réponses incohérentes pendant l'interview, tâches qui relevaient toutes du travail de bureau auparavant.

Les conséquences de l'utilisation de l'ITAO pour la collecte des données d'enquête pourraient être évaluées de plusieurs façons dont a déjà bien traité la littérature statistique. De nombreux articles ont décrit de nouvelles possibilités (ou options) de collecte des données qu'offre l'ITAO [11, 13, 17, 29, 41, 54]. D'autres articles ont étudié les effets de l'ITAO sur la qualité des données d'enquête, sur les estimations d'enquête et sur les coûts des enquêtes [7, 17, 29, 32]. Quelques articles ont examiné les effets de l'ITAO sur l'administration des enquêtes et sur la structure interne des organismes qui effectuent des enquêtes [2, 4, 9, 24]. Dans cet article on utilise une optique quelque peu différente; on y évalue l'effet de l'ITAO à partir de la situation et de la rapidité de la mise en application de l'ITAO dans les organismes gouvernementaux qui s'occupent de collecte de données ainsi que des conséquences de l'utilisation de l'ITAO et des obstacles qui y sont posés par ces organismes.

1.1 Utilisation courante et projetée de l'ITAO par les organismes du gouvernement américain

Quelle est l'importance de l'utilisation de l'ITAO par les organismes du gouvernement américain qui s'occupent de collecte de données? Le tableau 1 où l'on trouve le nombre de postes d'intervieweurs utilisables pour l'ITAO et d'emplacements (ou de lieux d'installation) d'ITAO pour quatre organismes fédéraux nous propose une réponse à cette question. Il se peut que l'on ait oublié de plus petits lieux d'installation utilisés par d'autres organismes américains. Dans ce tableau et dans ceux qui suivent, le nombre de postes d'ITAO comprend ceux qui sont utilisés par les surveillants et pour les gestionnaires ainsi que ceux qui servent directement à affectuer des interviews. Pour certains organismes, les mêmes postes servent aussi à la saisie des données à partir de formules imprimées et à d'autres fins.

Actuellement, les quatre organismes américains qui figurent au tableau 1 utilisent près de 500 postes d'ITAO. Si les plans actuels de ces organismes se réalisent comme prévu, alors, ainsi que le tableau 2 le montre, le nombre total de postes d'ITAO utilisés par ces organismes aura presque triplé en 1994. Un examen plus approfondi de leurs plans révèle des différences importantes entre les organismes pour ce qui est de l'utilisation qu'ils font de l'ITAO et de la rapidité avec laquelle ils mettent cette technique en application.

Actuellement, le Census Bureau exploite deux emplacements d'ITAO. Le premier est le centre téléphonique de Hagerstown de la division des opérations sur le terrain (Field Division), ce centre est surtout utilisé pour les enquêtes auprès des résidents dans les ménages et pour de petites enquêtes auprès des entreprises. Le nombre de postes d'ITAO qui se trouvent à cet emplacement passera de 32 à 72 en 1990, mais les plans visant à créer deux emplacements additionnels d'ITAO d'une importance comparable dans d'autres parties du pays d'ici 1994 dépendent de l'utilisation accrue de l'ITAO dans le cadre de l'enquête sur la population active, ce qui n'a pas encore été approuvé.

Le Census Bureau exploite aussi, par l'intermédiaire de sa division de la préparation des données (Data Preparation Division) à Jeffersonville, Indiana, un autre emplacement d'ITAO. À cet endroit, on recueille des données auprès d'établissements dans le cadre des enquêtes sur le commerce de gros et de détail de la division des entreprises (Business Division) du Census Bureau. Les 72 postes de travail qui se trouvent à cet emplacement sont des micro-ordinateurs interconnectés, ils servent à la réalisation d'interviews par la méthode de l'ITAO, à la saisie de données à partir de questionnaires reçus par la poste et au suivi des

questionnaires rejetés au contrôle pour ces mêmes enquêtes. Il n'existe pas de plans à court terme visant à accroître le nombre de postes de travail à cet emplacement, mais il faudra encore plusieurs années avant que les autres enquêtes de la division des entreprises soient effectuées par ITAO.

TABLEAU 1

EMPLACEMENTS ET POSTES D'ITAO UTILISÉS PAR DES ORGANISMES FÉDÉRAUX AMÉRICAINS, 1988

Organisme	Emplacements d'ITAO	Postes d'ITAO	Postes par emplacement
Census Bureau	2	104	52.0
Division des opérations sur le terrain	1	32	32.0
Division de la préparation des données	1	72	72.0
Bureau of Labor Statistics	14	62	4.4
Enquêtes sur l'indice des prix à la consommation	1	8	8.0
Enquêtes spéciales pour l'OEUS	1	18	18.0
Programme des statistiques courantes sur l'emploi	12	36	3.0
National Agricultural Statistics Service	14	193	13.8
National Centers for Disease Control	21	140	6.7
Total	51	499	9.8

Trois programmes de collecte de données du Bureau of Labor Statistics (BLS) utilisent l'ITAO. Le premier de ces programmes se déroule à un petit emplacement d'essai, à l'administration centrale, où l'on élabore des méthodes d'ITAO pour les enquêtes sur l'indice des prix à la consommation (IPC). En 1994, ce programme utilisera aussi un emplacement national comprennant de 40 à 60 postes d'ITAO où l'on effectuera des interviews dans un cadre d'exploitation. Le deuxième programme d'ITAO du BLS, utilise un emplacement de 18 postes de travail, à l'administration centrale, afin de réaliser des enquêtes spéciales pour l'Office of Employment and Unemployment Statistics du BLS. Le troisième et plus important essai d'ITAO effectué par le BLS se déroule dans le cadre du programme des statistiques courantes sur l'emploi (SCE). Dans cette enquête mensuelle auprès des établissements, l'ITAO est utilisée pour les interviews, pour le suivi dans les cas de non-réponse et pour la conciliation des questionnaires rejetés au contrôle. Si l'essai initial réussit, le BLS augmentera l'utilisation qu'il fait de l'ITAO dans le cadre du programme des SCE, de 12 bureaux d'État qui disposent de l'équipement nécessaire actuellement on passera à 51 bureaux en 1994.

Le National Agricultural Statistics Service (NASS) exploite actuellement le plus important réseau d'ITAO du gouvernement américain pour ses enquêtes relatives aux fermes et aux exploitants agricoles. Un programme d'approvisionnement présentement en cours comprendra une amélioration du matériel disponible dans les 14 bureaux d'État du NASS qui sont déjà équipés pour utiliser la méthode de l'ITAO et l'installation de matériel d'ITAO dans

TABLEAU 2

NOMBRE D'EMPLACEMENTS ET DE POSTES D'ITAO ACTUELS ET PROJETÉS DANS LES ORGANISMES FÉDÉRAUX AMÉRICAINS: 1988-1994

Organisme	(Emplacements) et postes d'ITAO			
	Actuels		Projetés	
	1988	1990	1992	1994
Census Bureau	(2)* 98	(2) 138	(3) 210	(4) 282
Division des opérations sur le terrain	(1) 32	(1) 72	(2) 144	(3) 216
Division de la prép. des données	(1) 72	(1) 72	(1) 72	(1) 72
Bureau of Labor Statistics	(14) 62	(14) 62	(32) 139	(53) 272
Enquêtes sur l'IPC	(1) 8	(1) 8	(1) 16	(1) 50
Enquêtes spéciales - pour l'OEUS	(1) 18	(1) 18	(1) 18	(1) 18
Programme des SCE	(12) 36	(12) 36	(30) 105	(51) 204
National Agricultural Statistics Service	(14) 193	(30) 375	(42) 747	(42) 747
National Centers for Disease Control	(21) 140	(21) 140	(26) 175	(26) 175
Total	(51) 499	(67) 721	(103) 1277	(125) 1482

* Le nombre d'emplacements d'ITAO est indiqué entre parenthèses.

tous les autres bureaux d'État, à l'exception de deux (Alaska et Hawaii) d'ici 1991. Un bon nombre des postes d'ITAO du NASS seront aussi utilisés pour la saisie des données à partir de questionnaires imprimés, pour le suivi dans les cas de questionnaires rejetés au contrôle et pour remplir des fonctions de bureautique.

Le National Centers for Disease Control (CDC) exploite un réseau de 21 bureaux d'État, équipés du matériel d'ITAO, où l'on effectue des interviews pour l'enquête de surveillance des facteurs de risque en matière de comportement et pour les enquêtes-ménages connexes effectuées par sondages téléphoniques à composition aléatoire (STCA) [37]. On ne prévoit qu'une croissance modeste du réseau d'ITAO du CDC au cours des prochaines années puisque, dans un bon nombre des autres États, la collecte des données est effectuée à contrat par des organismes d'enquête locaux.

On peut tirer trois conclusions générales à partir des plans de ces organismes:

Premièrement, les organismes du gouvernement américain qui augmentent actuellement leurs installations d'ITAO ou le nombre d'enquêtes pour lesquelles ils utilisent

cette méthode prévoient mener ces plans à terme entre 1990 et 1994. Tous ces plans n'auront pas l'ampleur prévue ou ne seront pas réalisés à temps, mais la fin de la phase initiale de mise en application de l'ITAO pour la collecte des données à l'échelle fédérale est en vue.

Deuxièmement, l'utilisation prédominante des installations internes d'ITAO par les organismes du gouvernement américain ne se fera pas pour appuyer des groupes considérables de personnes regroupées dans un bureau national et dans des bureaux régionaux afin d'effectuer des interviews téléphoniques. Plutôt, comme le tableau 3 le montre, l'ITAO sera habituellement utilisée pour appuyer des groupes moins considérables d'employés travaillant dans des bureaux au niveau des États et qui utilisent souvent le même matériel pour effectuer la saisie des données figurant sur les questionnaires reçus par la poste, la conciliation des questionnaires rejetés au contrôle et des activités connexes.

Troisièmement, même pour les organismes où les installations d'ITAO connaissent la croissance la plus rapide, cette méthode ne deviendra pas nécessairement le principal outil de collecte des données. Au NASS et pour le programme des statistiques courantes sur l'emploi du BLS les questionnaires expédiés par la poste demeurent des éléments extrêment importants des méthodes mixtes de collecte des données utilisées par ces organismes. Au Census Bureau, les 280 à 300 postes d'ITAO prévus pour 1994 représenteront toujours très peu de choses par rapport à un personnel sur le terrain composé de 2,500 à 3,000 recenseurs qui effectueront des interviews sur place. Pour le Census Bureau, le fait d'équiper son personnel sur le terrain de micro-ordinateurs portatifs afin de pouvoir appliquer la méthode d'interview sur place assistée par ordinateur (IPAO) représentera une tâche beaucoup plus considérable.

1.2 Facteurs qui influent sur la rapidité avec laquelle l'ITAO est mise en application

Puisque la fin de la période initiale de mise en application de l'ITAO est maintenant en vue pour les organismes du gouvernement américain, nous pouvons nous arrêter brièvement aux facteurs qui influent sur la rapidité avec laquelle cette méthode est mise en application. La première enquête, du secteur privé, pour laquelle on a utilisé l'ITAO, a été effectuée par la société Chilton Research en 1971; mais même les organismes gouvernementaux où les progrès sont les plus rapides n'appliqueront pas la méthode d'ITAO de façon intégrale à leurs principales enquêtes permanentes avant 1991. Vingt années, ou une génération entière, se seront alors écoulées depuis l'invention de l'ITAO. Dans le cas des principales enquêtes-ménages du Census Bureau, telles que l'enquête sur la population active et l'enquête nationale sur la criminalité, c'est en 1994, au plus tôt, que l'ITAO pourrait être appliquée de façon intégrale, mais il est possible qu'il y ait un certain retard dans les opérations et que le projet ne puisse être réalisé avant 1996. Dans un tel cas, un quart de siècle se serait écoulé depuis la réalisation, par le secteur privé, de la première enquête faisant appel à l'ITAO. Pourquoi la mise en application de cette méthode a-t-elle pris tant de temps?

De nombreux facteurs différents ont, sans aucun doute, contribué à ce que cette période de mise en application dure de 20 à 25 ans. Au début des années 70, d'après Dillman et Tarnai [9], les administrateurs des grosses enquêtes-ménages fédérales ont eu tendance à écarter l'interview téléphonique comme méthode inférieure de collecte de données. De plus, quand on était prêt à accepter un changement, il est arrivé souvent que la technologie nécessaire n'était pas disponible. C'est seulement à la fin des années 70, après leur élaboration par les organismes universitaires (avec l'aide du gouvernement) que les fonctions nécessaires pour de nombreuses enquêtes gouvernementales ont été intégrées au logiciel utilisé pour l'ITAO [40]. Après 1980, quand les organismes gouvernementaux ont commencé à jouer un rôle positif dans l'élaboration de l'ITAO à l'intérieur de leurs services, ils ont

TABLEAU 3

NOMBRE PROJETÉ D'EMPLACEMENTS D'ITAO, DE POSTES D'ITAO ET NOMBRE DE POSTES PAR EMPLACEMENT DANS LES ORGANISMES FÉDÉRAUX AMÉRICAINS, 1994

	Emplacements	Postes	Postes / Emplacement
Installations nationales et régionales	6	356	59.3
Census Bureau	4	288	70.5
BLS IPC & OEUS	2	68	34.0
Emplacements dans les États	119	1,126	9.5
NASS	42	747	17.8
BLS - SCE	51	204	4.0
CDC	26	175	6.7
Total pour les É.-U.	125	1,482	11.9

souvent entrepris des programmes de recherche considérables afin d'évaluer les effets de l'ITAO sur les coûts, sur la qualité des données et sur les estimations. Au U.S. Census Bureau ainsi qu'à Statistique Canada, ce n'est que maintenant que les rapports finals relatifs à ces programmes sont disponibles [7, 57]. La recherche effectuée a été particulièrement minutieuse dans les cas où un changement dans les méthodes de collecte de données aurait pu avoir un effet sur des séries chronologiques dont l'utilisation est délicate en matière de politiques. Comme les organismes du secteur privé qui effectuent des enquêtes, les organismes gouvernementaux ont appliqué plus rapidement l'ITAO aux nouvelles enquêtes ou à celles qui sont effectuées peu souvent. Les longs délais d'exécution et les retards fréquents dans la planification, la budgétisation et l'approvisionnement au niveau gouvernemental ont, sans aucun doute, encore ralenti la mise en application de l'ITAO par les gouvernements.

1.3 ITAO, STCA et centralisation

La période d'incubation prolongée de l'ITAO au niveau des gouvernements peut aussi être expliquée, en partie, par le fait qu'à l'origine elle a été associée à deux méthodologies connexes soit les sondages téléphoniques à composition aléatoire (STCA) et l'interview téléphonique centralisée [50, 59]. On utilise parfois l'expression "méthodes téléphoniques modernes" pour décrire collectivement la méthode des STCA, l'interview téléphonique centralisée et l'ITAO. Groves et Kahn [14] ont dit de l'évolution conjointe de ces méthodes qu'il s'agissait d'un des principaux progrès dans l'histoire des méthodes d'enquête, la classant avec le sondage aréolaire probabiliste et l'utilisation des ordinateurs pour analyser les données d'enquête. En 1980, la majorité des centres de recherche par sondage des universités américaines et des organismes d'étude de marché avaient établi des installations téléphoniques centralisées où l'on effectuait des STCA et d'autres enquêtes [4, 32, 48]. L'acquisition de l'équipement nécessaire pour appliquer la méthode d'ITAO a souvent suivi [49]. Les organismes gouvernementaux n'ont pas emboîté le pas.

Globalement les "méthodes téléphoniques modernes" ne pouvaient être appliquées facilement à la collecte de données effectuée par les gouvernements. C'est dans le cas de la méthode des sondages téléphoniques à composition aléatoire, dont la possibilité qu'elle offre

de réduire les coûts d'enquête a créé beaucoup d'intérêt parmi les statisticiens à l'emploi des gouvernements, que cette situation est la plus apparente. Le U.S. National Center for Health Statistics [53], le U.S. Census Bureau [3, 26, 45] et Statistique Canada [10] ont tous commencé leurs études relatives aux méthodes téléphoniques modernes par un essai prudent des sondages téléphoniques à composition aléatoire [27]. Cependant, comme Drew, Choudhry et Hunter [10] l'ont fait remarquer, peu d'enquêtes gouvernementales effectuées aux États-Unis ou ailleurs dans le monde utilisent la méthode des sondages téléphoniques à composition aléatoire. L'omission des ménages qui n'ont pas le téléphone (environ 7% du total aux États-Unis) ainsi que les taux de refus habituellement plus élevés dans le cas des interviews téléphoniques menées sans contact initial ont constitué des obstacles importants à l'utilisation de la méthode des STCA pour un bon nombre ou pour la majorité des enquêtes gouvernementales.

La méthode des sondages téléphoniques à composition aléatoire demeure une méthode d'échantillonnage précieuse dans le cas des populations avec un très grand nombre d'abonnés au téléphone et pour les enquêtes où les problèmes que cette méthode impose pour ce qui est du champ d'observation et de la non-réponse sont acceptables. Cependant, pour certains organismes statistiques gouvernementaux, l'accent mis au départ sur la méthode des STCA n'a servi qu'à détourner l'attention de ce qui semble maintenant être les utilisations plus fructueuses de l'ITAO. Ce n'est que lorsque la méthode des STCA a été éliminée comme méthode d'échantillonnage pour la majorité des enquêtes-ménages menées par le gouvernement américain, vers 1986 dans le cas du Census Bureau, que l'on a pu mettre en application des projets visant à utiliser l'ITAO dans des plans de sondage mixtes à base unique. L'adoption un peu plus rapide de l'ITAO dans le cas des enquêtes auprès des établissements peut être attribuable, en partie, au fait que, pour ces enquêtes, on s'en remet traditionnellement à l'échantillonnage sur liste. On n'a jamais songé à passer à la méthode des STCA . Le recours aux sondages téléphoniques à composition aléatoire n'a apparemment jamais été mis en question dans le cas de l'application de l'ITAO à l'enquête de surveillance des facteurs de risque en matière de comportement des Centers for Disease Control [37]. Dès le début de cette enquête, c'est la méthode des STCA qui a été choisie.

L'interview téléphonique centralisée constitue le deuxième élément important des méthodes téléphoniques modernes qui n'a pu être appliqué facilement à la collecte de données effectuée par les gouvernements. Dans les centres d'études de marché commerciaux et universitaires des États-Unis, le passage de la méthode où des intervieweurs locaux "dispersés" faisaient leurs appels à partir de leur propre domicile à une méthode où des intervieweurs "centralisés" font leurs interviews téléphoniques à partir de bureaux nationaux ou régionaux s'était fait, dans la majorité des cas, à la fin des années 70. Durant les années 80, seules les enquêtes-ménages gouvernementales continuaient de faire une utilisation importante des interviews téléphoniques dispersées réalisées à partir du domicile des intervieweurs.

L'interview téléphonique centralisée et l'ITAO possèdent non seulement des liens historiques, ces deux méthodes se renforcent aussi mutuellement. La société Chilton Research a été l'une des premières entreprises à avoir recours à l'interview téléphonique centralisée en 1966 [22] puis elle a introduit l'ITAO comme version automatisée de cette méthode cinq ans plus tard [29]. Il est plus facile d'offrir un soutien informatique à un groupe centralisé d'intervieweurs qui partagent le même matériel, les mêmes programmes, le même échantillon et qui traitent avec le même personnel technique [15, 16]. En même temps, l'ITAO favorise les interviews centralisées qui permettent de réaliser des gains d'efficacité et de profiter des fonctions que cette méthode peut offrir à un personnel nombreux comme l'ordonnancement automatique des appels, la surveillance en direct et la création des rapports relatifs aux opérations sur le terrain [9, 16, 30]. La centralisation des interviews contribue à la normalisation des procédures sur le terrain et au contrôle de la

qualité des interviews car elle facilite le recrutement, la formation et la surveillance des intervieweurs, alors que l'ITAO contribue à atteindre ces mêmes buts par l'adaptation des questions à la situation, par l'enchaînement informatisé des questions et par le contrôle en direct. La surveillance audiovisuelle du rendement des intervieweurs, qui ne peut être réalisée actuellement que dans le cas des intervieweurs qui utilisent l'ITAO de façon centralisée fournit les renseignements qui permettent d'assurer que les fonctions de l'ITAO permettant d'accroître la qualité des données sont utilisées de façon appropriée et que les intervieweurs dont le rendement n'est pas conforme aux normes sont reconnus et qu'ils suivent un complément de formation, au besoin.

L'interview téléphonique centralisée comporte aussi des désavantages potentiels. Si le rôle de l'intervieweur est trop spécialisé, cela peut diminuer le moral des intervieweurs [26]; de plus, les intervieweurs dans des bureaux centralisés n'ont pas les connaissances locales pas plus qu'ils ne possèdent l'accent et la compréhension des aspects culturels locaux qui sont tous des éléments que les intervieweurs locaux peuvent partager avec les répondants [24]. Quand, dans le cadre d'un plan de sondage mixte, on ne doit centraliser qu'une partie de la collecte des données d'enquête, il faut que l'efficacité de la gestion, de la coordination et des communications soit exceptionnelle [2]. Et, si la centralisation implique qu'il faut transférer le travail existant d'un groupe d'employés à un autre, on peut avoir à faire face à une résistance considérable de la part du personnel ou de l'organisme même [9]. Cela est particulièrement vrai dans le cas des enquêtes gouvernementales permanentes qui emploient fréquemment des intervieweurs, des surveillants et des gestionnaires comptant de nombreuses années de service.

Les problèmes les plus difficiles reliés à la mise en application de l'ITAO dans les organismes gouvernementaux semblent découler des questions d'ordre organisationnel relatives à l'interview centralisée par opposition à l'interview décentralisée que l'emploi de l'ITAO soulève généralement. Lyberg [24] signale, qu'au Bureau de la statistique de la Suède, la haute direction a réglé cette question au tout début en s'engageant à effectuer les interviews téléphoniques décentralisées à partir du domicile des intervieweurs tout en limitant l'ITAO centralisée à de petites enquêtes spécialisées et au suivi dans les cas de non-réponse. Les accords traditionnels entre organismes ou entre le gouvernement central et ceux des États (des provinces), les engagements officiels ou non envers les employés et les politiques prioritaires de l'administration au pouvoir peuvent se révéler des éléments aussi importants relativement à ces décisions que les buts de maximiser le contrôle des erreurs et l'efficience des opérations.

Dans le cas des enquêtes auprès des établissements, la question de la centralisation se présente sous une forme légèrement différente. Au début, les enquêtes de ce genre, menées par le gouvernement, se faisaient généralement par échange de correspondance à l'aide de méthodes d'envoi et de retour par la poste. Habituellement, on a commencé à effectuer des interviews téléphoniques dans le but d'inciter les répondants à poster leur questionnaire rempli, pour accepter au téléphone les réponses tardives et pour concilier les questionnaires rejetés au contrôle. Il se peut, afin de réaliser des gains d'efficacité, que les mêmes employés de bureau traitent les questionnaires retournés par la poste et effectuent les interviews téléphoniques. Les questions clés au niveau organisationnel sont de déterminer si les opérations de correspondance par la poste et/ou téléphonique devraient se tenir dans les bureaux national, régionaux ou des États (des provinces). L'apparition de l'ITAO a renforcé les arguments pour la centralisation. La Division des Entreprises du Census Bureau a pris cette orientation dans l'automatisation de ses enquêtes du commerce de gros et de détail. On a retiré les entrevues téléphoniques des 12 bureaux régionaux pour les centraliser au bureau de Jeffersonville de la division de la préparation des données où la mise à la poste, la saisie des données et le traitement initial étaient déjà effectués. Il a alors été plus facile d'introduire un système intégré combinant la mise à jour des bases de sondage, la mise à la

poste, l'ITAO, la saisie des données, le contrôle et les opérations relatives à la base de données tout comme cela a favorisé la formation d'un personnel centralisé pour employer ce système intégré ainsi que la supervision des employés qui l'utilisent. Comme ce changement a entraîné un transfert de charge de travail d'une division à une autre, il ne s'est pas accompli sans tensions occasionnelles et examen minutieux de la situation par l'administration, mais il ne fait aucun doute que la décision a facilité automatisation de ces enquêtes.

Quand des enquêtes sont trop considérables pour être effectuées à partir d'un seul emplacement national d'ITAO, c'est le recours aux bureaux régionaux qui constitue la meilleure solution de rechange. Le National Agricultural Statistics Service a examiné la possibilité d'utiliser une structure de ce genre pour ses enquêtes sur les fermes et sur les exploitants agricoles. Dans un des projets étudiés, chacun des 41 bureaux d'État serait responsable de ses propres interviews directes, mais seulement quelques bureaux d'État administreraient des installations d'ITAO d'où l'on effectuerait des interviews téléphoniques dans l'État où chacun de ces bureaux se trouve et dans les États voisins [2]. On a effectué un essai important de ce projet afin de déterminer si les bureaux du NASS pouvaient effectuer, de façon efficace, des interviews par ITAO à l'extérieur de l'État où chacun de ces bureaux se trouve. On a trouvé que ni le lieu où l'installation d'ITAO était située, ni le niveau d'expérience antérieur en ITAO avaient un effet sur la qualité des données ou sur les principales estimations [2]. Les résultats de la recherche entreprise n'ont permis d'éliminer aucun des emplacements des bureaux régionaux ou d'État pour l'implantation des installations d'ITAO du NASS. Les 41 bureaux d'État seront tous équipés pour l'ITAO bien que certaines enquêtes pourront être effectuées à partir d'un plus petit nombre de bureaux.

Il se peut que la centralisation des interviews téléphoniques puisse être plus importante dans le cas des enquêtes-ménages que dans celui des enquêtes auprès des établissements. Selon Werking, Tupek et Clayton [60] les erreurs de réponse dans les enquêtes-ménages proviennent surtout de sources comme les erreurs de mémoire et le télescopage. Il s'agit du genre de problèmes qui peuvent être réduits si l'on se conforme aux bons usages en matière d'interview, attitude qui est encouragée par la formation approfondie des intervieweurs ainsi que par leur surveillance et leur contrôle étroits. Par contraste, les principales sources d'erreur de réponse lors des enquêtes auprès des établissements sont habituellement dues aux systèmes d'archivage des établissements. La réduction de ces erreurs peut dépendre moins de la mesure dans laquelle les intervieweurs se conforment aux normes établies que de vérifications périodiques des dossiers afin de déterminer le genre exact de renseignements qui doivent être fournis par les établissements. Cet argument appuie ce qui semble maintenant être l'utilisation prédominante de l'ITAO par les organismes gouvernementaux américains, c'est-à-dire pour des enquêtes auprès des établissements effectuées à partir de bureaux d'État relativement petits et décentralisés.

Quels sont les principaux avantages qu'offre l'ITAO pour de telles enquêtes auprès des établissements? Tortora [54] et Werking et coll. [60, 61] ont mentionné des avantages connus comme: la commande de l'enchaînement des questions; le contrôle longitudinal ainsi que celui des gammes et de la cohérence en direct; la saisie et la collecte simultanées des données; des réductions dans le nombre de rappels à effectuer dans le cas des questionnaires rejetés au contrôle et l'amélioration de la gestion des enquêtes. Aucun des deux auteurs n'a soutenu que l'ITAO permettait de réduire le coût des enquêtes. Werking et ses collègues du BLS ont aussi mentionné que l'utilisation de l'ITAO à la place des méthodes de collecte par la poste dans le cadre du programme des statistiques courantes sur l'emploi permettait de réaliser des gains importants pour ce qui est d'améliorer les taux de réponse, de conserver un plus grand nombre de répondants lors des enquêtes à échantillon constant et d'améliorer l'actualité des données. Cependant, il se peut que ces derniers avantages découlent directement du fait que l'on a recueilli les données par téléphone plutôt que par la poste et non de l'emploi de la technologie de l'ITAO <u>en soi</u>.

1.4 Mise en application de l'ITAO pour les enquêtes-ménages à échantillon constant

La mise en application de l'ITAO pour les enquêtes-ménages gouvernementales pose souvent plus de problèmes au niveau de la conception et de l'organisation que ce n'est le cas pour les enquêtes auprès des établissements. Cette situation est illustrée par les plans du Census Bureau visant à introduire graduellement l'ITAO pour trois enquêtes-ménages nationales à échantillon constant: l'enquête sur la population active (EPA), l'enquête nationale sur la criminalité (ENC) et l'enquête sur le logement aux États-Unis (ELEU).

Chacune de ces enquêtes-ménages à échantillon constant utilise un plan de sondage mixte. Le premier contact avec chaque adresse dans l'échantillon se fait par visite sur place afin de déterminer les logements inadmissibles et pour encourager la participation des ménages. Le cinquième contact effectué dans le cadre de l'EPA et de l'ENC est aussi une visite sur place afin de rétablir un rapport personnel avec le ménage pendant que la série d'interviews est en cours. Les autres interviews sont effectuées par téléphone quand cela est possible et quand le répondant l'accepte ainsi que par interview sur place dans les autres cas.

Actuellement, les mêmes intervieweurs locaux effectuent les interviews sur place et les interviews téléphoniques, ces dernières étant effectuées depuis leur propre domicile. Comme l'EPA et l'ENC sont effectuées chaque mois, ces deux enquêtes emploient des intervieweurs permanents. L'ELEU nationale n'est effectuée que toutes les deux années et il faut reconstituer son personnel d'intervieweurs pour chaque volet de l'enquête. Il arrive souvent que les intervieweurs locaux travaillent, chaque mois, à plus d'une enquête du Census Bureau et ce personnel permanent constitue généralement le noyau des intervieweurs employés pour les enquêtes moins fréquentes. La surveillance des intervieweurs est assurée par 12 bureaux régionaux.

Quand l'ITAO sera introduite pour ces enquêtes, aucun changement ne sera apporté lors de la visite initiale à chaque adresse dans l'échantillon. Ces visites se feront encore sous forme d'interviews sur place puisqu'il est impossible, avec des interviews téléphoniques sans contact initial, d'obtenir des taux de réponse comparables et de conserver un aussi grand nombre de participants à des enquêtes à échantillon constant [26]. L'ITAO remplacera les interviews téléphoniques dispersées effectuées depuis le domicile des intervieweurs locaux à partir de la deuxième visite lors des enquêtes à échantillon constant. Cette façon d'effectuer le travail sur le terrain présente de nombreux avantages potentiels: (1) réduction des coûts sur le terrain, (2) réduction des problèmes liés au recrutement des intervieweurs sur les marchés où il y a pénurie de main-d'oeuvre et (3) amélioration possible des estimations d'enquête comme cela est décrit à la section suivante [21, 34, 57]. Néanmoins, la transition d'une méthode à l'autre pose un certain nombre de problèmes de conception et d'organisation à la résolution desquels il faut consacrer du temps et des efforts.

Le premier de ces problèmes est celui qui porte sur l'élaboration de méthodes appropriées pour effectuer le transfert rapide, mais contrôlé, de chaque dossier entre les modes d'interview sur place et d'ITAO. Quand l'interview sur place effectuée lors de la première visite est terminée, les données sur le dénombrement du ménage ainsi que les données contenues dans les dossiers relatifs au travail sur le terrain doivent être introduites dans un ordinateur et transformées en fichiers informatiques pour les deuxième et dernière interviews qui seront effectuées par ITAO. Les dossiers doivent aussi être transférés de l'ITAO, effectuée aux bureaux régionaux, aux intervieweurs locaux dans le cas des cinquièmes interviews effectuées dans le cadre de l'EPA et de l'ENC ainsi que pour le suivi sur place dans le cas des ménages qui ne peuvent être atteints par la méthode de l'ITAO. Des procédures ont été élaborées afin d'accomplir ces transferts dans les délais prévus, mais le processus qui consiste à faire un transfert répété des dossiers entre le personnel qui

s'occupe des interviews dispersées et celui qui s'occupe des interviews centralisées (ainsi qu'entre les dossiers sur papier et les fichiers informatiques) est complexe et sujet à erreur. Le passage à un système intégré d'ITAO-IPAO, qui utiliserait, sur le terrain, l'interview sur place assistée par ordinateur faciliterait des transferts plus rapides et plus fiables.

Le second problème de transition lors de l'introduction de l'ITAO pour les enquêtes-ménages à échantillon constant est celui qui touche à la réduction temporaire de l'efficacité des plans de sondage. L'EPA ainsi que l'ENC emploient toutes deux des échantillons par grappes choisis principalement afin de minimiser le coût des tâches d'interview qui comprennent à la fois des interviews sur place et des interviews téléphoniques dispersées. Quand les interviews téléphoniques dispersées sont transférées à l'ITAO, il se peut que les visites sur place qui restent à effectuer ne constituent plus des tâches sur le terrain acceptables ou efficaces. Dans les unités primaires d'échantillonnage (UPÉ) confiées à un seul intervieweur, on ne peut tout simplement pas transférer les interviews téléphoniques à l'ITAO sans amener la charge de travail des intervieweurs (et leur rémunération) en-dessous des niveaux acceptables. Dans le cas des UPÉ confiées à plusieurs intervieweurs, on peut redistribuer les interviews sur place une fois que l'on a enlevé les cas réglés par téléphone, mais comme il s'agit de fragments de grappes antérieures, il se peut que ces interviews soient mal réparties pour qu'un plus petit nombre d'intervieweurs puissent se déplacer de façon efficace. La solution évidente consiste à utiliser un plan de sondage révisé qui donne des affectations sur le terrain efficaces après transfert de la majorité des interviews téléphoniques à l'ITAO, mais les échantillons utilisés pour l'EPA et pour l'ENC sont basés sur le recensement décennal et ils ne sont révisés qu'à tous les dix ans. Un nouveau plan de sondage ne peut donc être pleinement appliqué avant 1995.

Le troisième problème relié au passage des interviews téléphoniques dispersées à l'ITAO est celui de la réduction du personnel sur le terrain et de l'augmentation simultanée de celui travaillant à l'ITAO. Dans le cas de l'EPA, la plus considérable des enquêtes courantes du Census Bureau, la transition se fera, au début, à la suite du départ des intervieweurs sur le terrain [57]. Les dossiers ne seront transférés à l'ITAO qu'à mesure que les intervieweurs qui travaillent à l'EPA sur le terrain quitteront le Census Bureau pour d'autres raisons. La rapidité et l'importance de la mise en application de l'ITAO seront donc limitées, au début, par le taux de départ des intervieweurs travaillant dans des UPÉ confiées à plusieurs intervieweurs.

Le quatrième et dernier problème de transition est celui qui consiste à trouver un volume de travail suffisant pour les intervieweurs qui utilisent l'ITAO. Les interviews réalisées dans le cadre de l'EPA sont effectuées pendant la troisième semaine de chaque mois et la majorité de celles relatives à l'ENC pendant la première semaine du mois avec un certain nombre pendant la deuxième. L'interview centralisée par la méthode d'ITAO est limitée à un nombre encore moins considérable de jours par mois afin de permettre le suivi sur place des cas qui ne peuvent être réglés par téléphone. Ces deux enquêtes fourniront relativement peu de jours d'emploi chaque mois au personnel qui travaille à l'ITAO. C'est dans le cas de l'EPA que ce problème est le plus grave. Des trois installations d'ITAO dont le Census Bureau prévoit disposer en 1994, deux seraient presque exclusivement employées à l'EPA. Il se peut qu'il soit nécessaire de prévoir des enquêtes additionnelles effectuées par ITAO afin de fournir plus d'heures de travail, ou qu'on doive modifier la conception du travail effectué sur le terrain dans le cadre de l'EPA ou limiter le nombre d'interviews de l'EPA transmises à l'ITAO si l'on veut utiliser, de façon économiquement viable, l'ITAO en mode mixte dans le cadre de l'EPA.

Des quatre problèmes de transition qui viennent d'être mentionnés, seul le premier découle de l'utilisation de la technologie de l'ITAO. Le transfert des dossiers entre les intervieweurs locaux dispersés et les intervieweurs centralisés qui utilisent l'ITAO est

compliqué par le passage des dossiers traités manuellement aux fichiers informatisés. Les problèmes d'efficacité de l'échantillonnage effectué sur le terrain, de la diminution du personnel travaillant sur le terrain et du manque de travail dans les installations d'ITAO découlent de la centralisation d'interviews dispersées auparavant. Ces problèmes seraient identiques, que l'installation centralisée utilise la méthode de l'ITAO ou des méthodes manuelles. Les principaux problèmes reliés à la mise en application de l'ITAO dans le cadre des enquêtes-ménages effectuées par les gouvernements découlent non de ses exigences technologiques, mais de ses exigences organisationnelles.

1.5 Quelques résultats récents

Bien que l'introduction de l'ITAO (et de l'interview téléphonique centralisée) pour les enquêtes-ménages permanentes effectuées par les gouvernements présente un certain nombre de problèmes au niveau de la conception et de l'organisation, il existe des preuves croissantes que ces méthodes peuvent entraîner des avantages importants, même s'ils sont parfois inquiétants. Les exemples présentés ici sont tirés des études d'évaluation de l'ITAO dans des plans de sondage mixtes effectuées par le Census Bureau. Ils diffèrent, sous deux rapports importants, des évaluations de l'ITAO faites à Statistique Canada par Catlin, Ingram et Hunter et font l'objet de la communication suivante lors du présent symposium [7].

Tout d'abord, la méthode de recherche utilisée par Catlin, Ingram et Hunter, compare l'ITAO centralisée aux interviews téléphoniques manuelles centralisées. On peut généralement attribuer les différences observées à l'emploi de l'ITAO plutôt qu'à celui des méthodes manuelles. Les études effectuées par le Census Bureau comparent la collecte des données en mode mixte qui utilise l'ITAO centralisée à la collecte des données en mode mixte qui utilise la méthode manuelle de l'interview téléphonique dispersée. Ainsi, on ne peut facilement distinguer les effets dus à l'emploi de l'ITAO de ceux qui sont dus à l'emploi de la méthode d'interview téléphonique centralisée.

Deuxièmement, Catlin, Ingram et Hunter ont surtout fait porter leur analyse sur les mesures transversales de la qualité des données telles que les taux de réponse par unité et par question, la cohérence des questions dans la même interview et la possibilité de coder les inscriptions relatives à la profession et à l'activité économique. Les résultats du Census Bureau présentés ici se concentrent sur les estimations établies à partir des résultats d'enquêtes à échantillon constant.

Schwanz, Montfort et Cannon [42] ont examiné l'utilisation de l'ITAO dans la composante nationale de l'enquête sur le logement aux États-Unis (ELEU), qui est menée tous les deux ans, lors des années impaires. En 1987, deux des six échantillons constants de l'ELEU ont été traités par ITAO afin d'évaluer l'utilité de cette méthode dans les interviews qui utilisent des échantillons constants avec des types sélectionnés de logements déjà interviewés. Bien que 39.5% des logements qui avaient été affectés à l'ITAO n'aient pu être atteints par téléphone et aient dû être recyclés afin que les interviews soient effectuées sur le terrain, les auteurs de l'étude ont conclu que les premières tentatives d'ITAO ont permis de réaliser des économies de temps et d'argent. Ils recommandent une utilisation semblable de l'ITAO dans les six échantillons constants nationaux de l'ELEU en 1989.

Cette étude a aussi porté sur la valeur des interviews de conciliation assistées par ordinateur pour ce qui est d'estimer les changements dans les caractéristiques des logements. Quand l'interview a été effectuée par ITAO, les réponses fournies en 1987 pour neuf questions principales ont été comparées aux réponses fournies par le même ménage pour ces questions en 1985. Les valeurs de 1985 n'étaient pas accessibles à l'intervieweur préposé à l'ITAO avant que la dernière réponse à l'interview de 1987 n'ait été introduite. Quand les différences dans les réponses fournies en 1985 et en 1987 dépassaient les limites

de tolérance prévues dans le programme, une question de conciliation était affichée afin que l'intervieweur demande au répondant de justifier la discordance. Cette technique est bien connue dans le cas des enquêtes auprès des établissements, où on l'appelle "contrôle longitudinal" ou "utilisation de données chronologiques" [35]. Elle est moins connue dans le cas des enquêtes-ménages où elle est connue sous le nom d'interview "avec comparaison" ou d'interview "avec rapprochement".

Plus de la moitié des interviews de l'ELEU réalisées par ITAO (54.8%) ont été rejetées à au moins un des contrôles longitudinaux et ont entraîné l'affichage d'au moins une question de rapprochement. Les réponses aux questions de rapprochement laissent supposer qu'environ la moitié des différences relevées correspondaient à des changements réels dans les caractéristiques du logement et environ la moitié étaient des erreurs de réponse. Si les changements dans les caractéristiques du logement étaient estimés à partir de tableaux croisés conventionnels des réponses fournies en 1985 et en 1987 aux mêmes questions pour les mêmes logements, près de la moitié des changements estimés seraient illusoires.

Le tableau 4 présente des exemples. Lors des interviews par ITAO, 19.2% des ménages ont déclaré, en 1987, un type de système de chauffage différent de celui déclaré en 1985. Cependant, quand on leur a signalé la différence dans les réponses, le changement déclaré a été ramené à 5.7% alors que le répondant a attribué le 13.5% qui reste à des erreurs soit dans la déclaration de 1985, soit dans celle de 1987. De même, alors que 12.0% des ménages ont déclaré, en 1987, un genre de sous-sol différent de celui déclaré en 1985, il semble que les changements réels n'aient été que de 1.6% alors que le 10.4% qui reste dans les différences entre les deux déclarations était dû à des erreurs de déclaration. Curieusement, les erreurs dans les réponses que le répondant a fournies en 1987, réponses qui avaient été données seulement quelques minutes auparavant, ont été mentionnées aussi souvent que les erreurs dans les réponses fournies, en 1985, soit deux ans auparavant. Il se peut qu'un tel contrôle longitudinal se révèle essentiel pour estimer les changements, au cours des ans, dans les caractéristiques non seulement des logements mais aussi des membres du ménage.

TABLEAU 4

SOURCES DÉCLARÉES DES DIFFÉRENCES ENTRE DES CARACTÉRISTIQUES CHOISIES DES LOGEMENTS EN 1985 ET EN 1987
ÉCHANTILLON TRAITÉ PAR ITAO DE L'ENQUÊTE SUR LE LOGEMENT AUX ÉTATS-UNIS

Question relative	Différence au logement totale 1985-1987	Rapprochement	
		Changement apporté	Erreur déclarée
Système de chauffage	19.2%	5.7%	13.5%
Nombre de chambres à coucher	12.3	7.6	4.6
Genre de sous-sol	12.0	1.6	10.4
Valeur de la maison	10.2	7.7	2.5
Combustible utilisé	9.9	5.2	4.6
Nombre de salles de bains	8.3	1.9	6.4

Bien que le contrôle longitudinal soit facilité à la fois par l'ITAO et par l'interview téléphonique centralisée, il ne fait aucun doute que ce contrôle pourrait aussi être effectué si l'on utilisait la méthode de l'interview sur place assistée par ordinateur (IPAO) dispersée sur le terrain. Pafford et Coulter [35] ont recommandé avec insistance de ne pas utiliser de données chronologiques dans les interviews manuelles où la connaissance qu'a déjà l'intervieweur des réponses fournies auparavant peut entraîner des biais dans les réponses ainsi qu'amener l'intervieweur à varier l'utilisation qu'il fait de ces renseignements.

Hubble et Wilder [21] ont analysé les effets de l'ITAO (et (ou) de l'interview téléphonique centralisée) sur les estimations découlant des données recueillies lors de l'enquête nationale sur la criminalité (ENC). Cette enquête porte sur un échantillon constant, avec renouvellement, d'environ 10,000 ménages par mois qui sont interviewés tous les six mois. On pose à chaque personne, âgée de douze ans et plus, vivant dans un ménage de l'échantillon des questions de sélection portant sur les actes criminels commis contre sa personne au cours des six mois précédant le mois de l'interview. On pose aussi au premier répondant du ménage des questions de sélection portant sur les actes criminels commis contre le ménage pendant la même période. Si un répondant répond "Oui" à une quelconque des questions de sélection, l'intervieweur remplit un rapport d'acte criminel pour chaque acte. Si le répondant ne répond jamais "Oui" aux questions de sélection, l'interview est très brève.

Les interviews téléphoniques de l'ENC seront graduellement effectuées par ITAO à commencer par les UPÉ confiées à plusieurs intervieweurs dans les régions difficiles à dénombrer. Le plan de sondage permet de comparer les résultats obtenus par ITAO avec ceux d'un groupe témoin pour lequel on continue d'effectuer des interviews téléphoniques dispersées. En 1987, quand les chiffres qui figurent au tableau 5 ont été obtenus, on affectait de 200 à 400 ménages par mois au groupe expérimental, celui sur lequel portait l'essai de l'ITAO. Toutes les interviews relatives au groupe expérimental ne se sont pas faites par ITAO. Il fallait se rendre sur place dans le cas de la première et de la cinquième interview semestrielles faites auprès de chaque ménage dans l'échantillon. (Les données de la première visite ne servent qu'à établir des limites et les estimations sont basées sur les données obtenues lors de la deuxième à la septième interview.) Il y a aussi des interviews sur place pour des ménages dans l'échantillon au cours d'autres mois dans le cas des ménages sans téléphone, pour ceux où une interview téléphonique est inacceptable ou inappropriée et dans le cas des ménages qui ont été recyclés pour faire l'objet d'une interview sur place quand on ne peut les atteindre par ITAO.

Un objectif important de la recherche en ITAO effectuée dans le cadre de l'ENC vise à évaluer les effets de l'utilisation de l'ITAO, en mode mixte, sur les taux d'actes criminels déclarés lors de l'enquête. Puisque l'on croit généralement que les actes criminels font l'objet d'une sous-déclaration, on suppose habituellement que des taux de déclaration plus élevés sont meilleurs. Comme le tableau 5 le montre, la méthode d'essai qui utilise l'ITAO a permis d'obtenir des taux de déclaration significativement plus élevés tant dans le cas des actes criminels commis contre la personne que de ceux commis contre le ménage. Le taux d'actes criminels commis contre la personne était de 29% plus élevé dans le cas du groupe expérimental que dans celui du groupe témoin. Pour les actes criminel commis contre le ménage, le taux du groupe expérimental était de 13% plus élevé.

Les interviews téléphoniques ont été la source de la différence globale. Il n'y avait pas de différence significative entre les taux d'actes criminels pour le groupe expérimental et pour le groupe témoin dans le cas des interviews sur place; mais dans le cas des interviews téléphoniques on a obtenu, pour le groupe expérimental, des taux de plus de 50% plus élevés que pour le groupe témoin dans le cas des actes criminels commis contre la personne. Bien que cela ne soit pas montré dans des tableaux distincts, l'effet n'est pas limité aux actes criminels moins sérieux ou à ceux qui constituent des cas limites, il semble être encore plus marqué dans le cas des actes criminels qui n'ont pas été déclarés à la police.

TABLEAU 5

TAUX ESTIMÉS D'ACTES CRIMINELS PAR 1,000 HABITANTS
DANS L'EXPÉRIENCE D'ITAO EFFECTUÉE DANS LE CADRE DE L'ENC, 1987

Genre d'actes criminels et mode d'interview	Taux d'actes criminels par 1,000 habitants		
	Groupe expérimental(E) avec ITAO	Groupe témoin(T)	Rapport E/T
Actes criminels contre la personne			
Total	113.4	87.9	1.29**
Interview sur place	108.6	116.4	0.93
Interview téléphonique	115.5	73.9	1.56**
Actes criminels contre le ménage			
Total	194.8	172.4	1.13*
Interview sur place	185.8	194.8	0.95
Interview téléphonique	198.9	161.1	1.23*

* Statistiquement significatif au niveau .10.
** Statistiquement significatif au niveau .05.

Ce résultat n'est pas nécessairement attribuable à l'ITAO. Il peut être dû aux occasions qu'offre l'interview téléphonique centralisée pour ce qui est de la surveillance et du contrôle direct et (au besoin) du recyclage des intervieweurs. Il se peut que les intervieweurs sur le terrain, sans surveillance, qui appellent de leur propre domicile ne se conforment pas autant aux procédures d'enquête précisées. Ou, il se peut que l'ITAO ainsi que l'interview téléphonique centralisée contribuent à l'effet relevé.

Les causes de cet effet ne présentent pas seulement un intérêt théorique. Si le soutien informatique permet d'augmenter le nombre de déclarations d'actes criminels, alors les intervieweurs sur le terrain, qui utilisent du matériel d'IPAO et appellent de leur propre domicile, devraient obtenir d'aussi bons résultats. Si c'est l'interview téléphonique centralisée qui est la cause principale de l'augmentation du nombre de déclarations, alors des intervieweurs téléphoniques centralisés qui utilisent des méthodes manuelles devraient faire aussi bien. Le choix le plus sûr est d'utiliser les deux méthodes. Le Census Bureau et le Bureau of Justice Statistics ont accepté de transférer à la méthode d'ITAO, d'ici 1991, autant d'interviews téléphoniques relatives à l'ENC que le plan de sondage actuellement utilisé pour cette enquête le permet. Les essais futurs de l'ITAO dans le cadre de l'ENC pourront nous en apprendre plus à propos de l'importance relative du soutien informatique et de l'interview centralisée sur cet effet.

Bushery et Cahoon [6] ont fourni un troisième résultat intéressant dans l'analyse qu'ils ont effectuée des effets de la collecte des données par ITAO, en mode mixte, sur des estimations relatives à la population active provenant des données recueillies lors de

l'enquête sur la population active. Bien que ce résultat ne soit pas aussi spectaculaire que l'effet apparent de l'utilisation de l'ITAO sur les taux d'actes criminels, il pourrait se révéler généralement plus pertinent s'il était confirmé par d'autres recherches.

La tendance connue sous le nom de biais dû au groupe de renouvellement a été relevée pour la première fois dans l'enquête sur la population active en 1955. Bailar [1] mentionne que ce phénomène est courant dans les enquêtes à échantillon constant: plus les répondants qui font partie d'un échantillon constant sont exposés à une enquête, moins la probabilité qu'ils déclareront un bon nombre des renseignements, que l'enquête a été conçue pour mesurer, est élevée. Dans l'EPA, la tendance générale n'a pas varié beaucoup depuis 1970. Les taux d'emploi, de chômage et d'activité diminuent du premier au troisième mois au cours desquels les répondants font partie de l'échantillon. Une question additionnelle posée pour la première fois lors du quatrième mois de l'enquête provoque une légère amélioration de ces statistiques pour ce mois. La tendance se dessine à nouveau, à partir d'un niveau inférieur, quand les mêmes ménages reviennent, un an plus tard, pour participer à un second ensemble de quatre interviews.

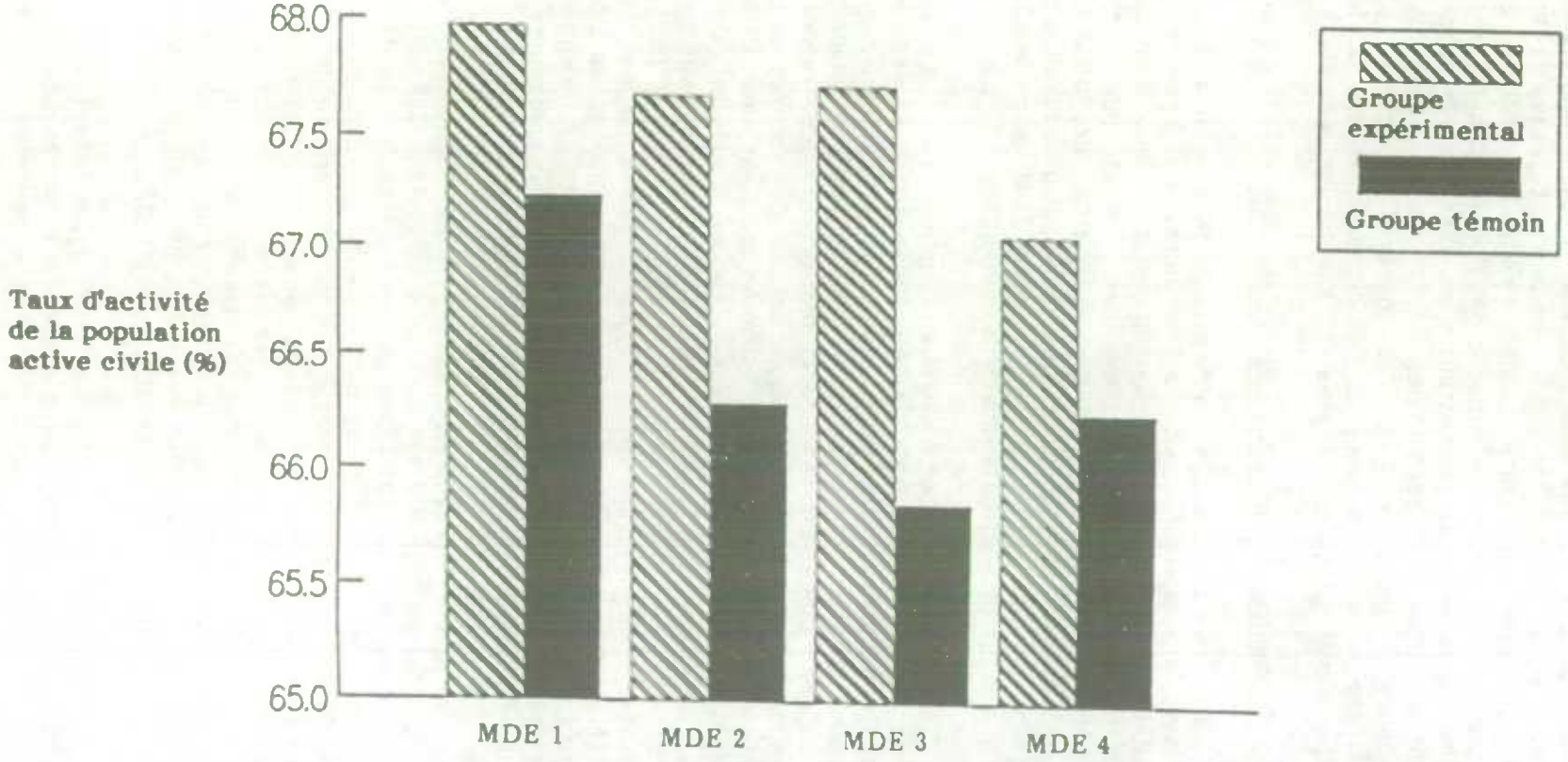
Dans la figure 1, on compare le taux d'activité de la population active civile pour l'EPA dans les quatre premiers mois au cours desquels les répondants font partie de l'échantillon pour (1) un groupe expérimental pour lequel on utilise l'ITAO pour la majorité des interviews téléphoniques au cours des deuxième, troisième et quatrième mois pendant lesquels les répondants font partie de l'échantillon et (2) un groupe témoin pour lequel on utilise surtout l'interview téléphonique dispersée pendant les mêmes mois. Dans les deux groupes, on a eu recours à des interviews sur place lors du premier mois où les répondants font partie de l'échantillon ainsi que dans le cas des ménages qui n'ont pas le téléphone et pour d'autres ménages spéciaux pendant les deuxième, troisième et quatrième mois. Le groupe témoin illustre la tendance classique des groupes de renouvellement de l'EPA alors que la tendance que présente le groupe expérimental a été considérablement différente. Quand l'ITAO remplace l'interview téléphonique dispersée, le biais caractéristique dû au groupe de renouvellement de l'EPA semble diminuer.

Nous ne nous arrêterons pas à étudier d'autres différences relevées dans ces résultats préliminaires, entre le groupe expérimental et le groupe témoin, comme les différences dans le niveau ainsi que les différences selon le mois de la présence dans l'échantillon. Mais si dans des recherches additionnelles on continue de trouver que l'ITAO (ou l'interview centralisée) diminue le biais dû au groupe de renouvellement, cela peut laisser supposer que ce phénomène bien connu dépend, du moins en partie, des méthodes d'interview sur le terrain plutôt que de n'être qu'une conséquence des caractéristiques des répondants qui changent quand on leur pose plusieurs fois les mêmes questions d'enquête.

Encore une fois, nous devons faire preuve de prudence pour ce qui est d'attribuer cette conclusion apparente à l'ITAO. Elle peut être une conséquence de la surveillance étroite du travail des intervieweurs qui est possible quand on a recours à la méthode de l'interview téléphonique centralisée. Dans l'un ou l'autre cas, les résultats tirés de l'ELEU, de l'ENC et de l'EPA peuvent nous avoir fait découvrir de nouvelles possibilités qui méritent une étude sérieuse. L'interview téléphonique assistée par ordinateur ainsi que l'interview téléphonique centralisée peuvent ne pas être seulement des commodités technologiques facilitant la collecte courante des données d'enquête. Ces méthodes peuvent se révéler des innovations importantes qui changent notre optique de la nature de la collecte des données d'enquête.

FIGURE 1

Taux moyen d'activité de la population active civile novembre 1986 à mars 1988



2. Technologies de la phase 2

Les technologies de la phase 2 sont celles dont la première utilisation importante pour la collecte des données d'enquête s'est produite après 1980 ou dont une telle utilisation pourrait se faire avant l'an 2000. Ces technologies comprennent des produits découlant de la révolution micro-informatique, mais n'y sont pas limitées. Puisque, jusqu'ici, on connaît peu ou pas leurs effets sur la collecte des données d'enquête, nous ne pouvons que décrire les technologies les plus prometteuses et spéculer sur leurs effets probables. Dans cette communication les technologies de la phase 2 sont divisées en quatre domaines: (1) technologies utilisables pour les enquêtes auprès des établissements, (2) technologies relatives au positionnement et aux communications, (3) interview sur place assistée par ordinateur et méthodes connexes et (4) technologie vocale.

2.1 Technologies utilisables pour les enquêtes auprès des établissements

Bien que les nouvelles technologies de collecte des données (celles qui suivent l'ITAO) puissent être appliquées aux enquêtes-ménages et aux enquêtes auprès des établissements, il semble que leur introduction soit un peu plus facile dans le cas de ces dernières. Il arrive souvent que les enquêtes auprès des établissements n'exigent que de brèves réponses numériques fournies par les mêmes répondants à des intervalles fréquents. Il se peut que ces répondants experts accueillent favorablement de nouvelles méthodes de collecte de données qui utilisent une technologie bien connue dans les entreprises, particulièrement quand ces méthodes minimisent le temps qu'ils doivent consacrer à ces enquêtes.

Le Bureau of Labor Statistics expérimente avec succès la simulation de la parole et la saisie des données à l'aide de téléphones à clavier dans le cadre de son programme des statistiques courantes sur l'emploi [60, 61]. Ce système ressemble à ceux que de nombreuses banques utilisent maintenant pour les transactions téléphoniques portant sur des comptes bancaires. Quand les répondants ont préparé leur rapport mensuel relatif aux SCE, ils composent un numéro de téléphone qui leur permet d'accéder à un bureau local du BLS où ils sont accueillis par une voie simulée puis on leur demande d'introduire leur numéro d'identification, à l'aide de leur téléphone à clavier. On leur pose ensuite une série de questions auxquelles ils doivent répondre en introduisant des données numériques à l'aide de leur téléphone, la confirmation de ces réponses est faite, au besoin, à l'aide d'une voie simulée. Puisque l'appareil fonctionne 24 heures par jour, sans la présence d'un intervieweur au téléphone, les coûts sont minimes. Bien entendu, un intervieweur du BLS téléphone aux répondants qui n'ont pas appelé à la date limite. Au cours de la phase suivante de ce projet, on étudiera l'utilisation de modules de reconnaissance de la parole pour remplacer l'introduction des données à l'aide de téléphones à clavier.

La Energy Information Agency (EIA) étudie une autre méthode, elle emploie des questionnaires informatisés à remplir soi-même [23] pour son système de déclaration des réserves de pétrole [52]. Des visites à un petit échantillon de répondants ont permis de constater que la majorité de ces derniers utilisaient déjà un ordinateur personnel soit dans leur bureau, soit dans un bureau voisin. On a préparé pour chaque société des disquettes sur lesquelles sont enregistrés: (1) un questionnaire informatisé à remplir soi-même auquel il faut répondre chaque mois, (2) des procédures à base de menus afin d'aider le répondant à obtenir les renseignements nécessaires à partir d'autres fichiers ou d'ordinateurs interconnectés et (3) des procédures programmées afin de composer le numéro des bureaux de l'EIA quand le questionnaire est rempli, d'introduire le mot de passe approprié et de transmettre les données à l'ordinateur de l'EIA.

Les organismes statistiques ne sont pas les seuls à élaborer de nouvelles méthodes de transmission des données dans le cas des enquêtes auprès des établissements. Au moins une nouvelle méthode a apparemment été inventée par les répondants: l'utilisation d'un

télécopieur pour transmettre une copie du questionnaire imprimé à l'organisme de collecte des données. Il se peut que les grandes entreprises trouvent commode de transmettre de gros fichiers sur des bandes pour ordinateur.

Il ne fait aucun doute qu'à mesure que la technologie informatique et que celle des communications évolueront, d'autres méthodes de collecte et de transmission des données seront découvertes. À long terme, il se pourrait que la meilleure stratégie de collecte de données dans le cas des enquêtes auprès des établissements soit un empressement à accepter la combinaison de méthodes que les répondants trouvent la plus commode. La méthode de l'ITAO pourrait n'être employée que comme méthode de réserve dans le cas des non-réponses, des déclarations produites en retard et des réponses incohérentes.

2.2 Systèmes de positionnement et de communication sur le terrain

C'est aux interviews téléphoniques que les progrès technologiques réalisés au cours des années 70 ont le plus contribué. Il semble probable que les progrès technologiques déjà réalisés et qui le seront au cours des années 80 et 90 contribueront le plus aux interviews sur place. Ces progrès peuvent comprendre des technologies permettant de saisir les réponses fournies lors des interviews sur place et des technologies permettant d'augmenter le genre de données recueillies.

Lors de la planification du recensement de 1990, le Census Bureau a examiné un certain nombre de systèmes de positionnement ou de navigation afin de déterminer si l'on pouvait disposer d'un matériel économique et pouvant être installé dans une automobile, qui fournirait des coordonnées géographiques sûres, avec une bonne répétabilité et une erreur inférieure à 30 pieds (ou approximativement 10 mètres). Ce matériel pourrait servir à de nombreux usages lors des recensements et des enquêtes, par exemple: (1) mise à jour des cartes, il pourrait être possible d'ajouter une rue sur une carte accessible en direct en parcourant cette rue, (2) enregistrement des coordonnées précises des constructions dans les régions rurales afin de faciliter les visites de rappel sur le terrain, (3) géocodage de l'emplacement précis des logements afin de permettre de grouper et de totaliser les données du recensement en fonction de toute région géographique dont les limites peuvent être définies à l'aide de coordonnées et (4) utilisation, à des fins d'échantillonnage, de régions géographiques aussi flexibles, qui ne seraient pas modifiées par les changements dans les routes, dans les lignes de transport d'énergie ainsi que les limites politiques.

On a trouvé et examiné de près certains systèmes dont les caractéristiques s'approchaient de celles qui étaient recherchées. Ces systèmes comprennent:

Le LORAN-C (acronyme de "long range navigation" (système de navigation à longue portée)) utilise des signaux radio à basse fréquence afin de déterminer une position sur la surface du globe [18]. C'est en navigation maritime que l'on utilise surtout le LORAN-C, mais des appareils qui peuvent être installés dans une automobile sont disponibles pour des usages terrestres. Malheureusement, ils ne se sont pas révélés suffisamment précis et fiables pour répondre aux besoins du Census Bureau quand on les a évalués pour la dernière fois en 1983. De plus, le système LORAN-C ne peut être utilisé en Alaska, à Hawaii et dans le centre des États-Unis.

Le système de positionnement global ("GPS") NAVSTAR est un système de radionavigation par satellite élaboré pour le département de la défense des États-Unis. Quand tout le système sera disponible, il permettra d'obtenir un positionnement en trois dimensions partout sur la terre [25]. Bien qu'il paraît que le "service de positionnement précis" du NAVSTAR aura une précision de 16 mètres, son utilisation sera réservée à des

usages militaires. Un "service de positionnement ordinaire" qui sera offert à un plus grand nombre d'utilisateurs aura une précision de 25 à 30 mètres. Le système de positionnement global NAVSTAR aurait dû être opérationnel en 1989, mais sa mise en service a été retardée à cause de la longue interruption des envolées de la navette spatiale.

L'ETAK (il s'agit d'un mot polynésien plutôt que d'un acronyme) est un système de navigation qui s'installe dans une automobile et qui incorpore un compas magnétique numérique, des détecteurs de mouvement ainsi qu'un système "intelligent" de navigation à l'estime [47]. Ce système indique la position de l'automobile sur une carte qui se déplace sur un écran de visualisation placé près du conducteur. Il faut disposer de cartes numérisées, enregistrées sur bandes magnétiques, de la région où l'on se trouve et les cassettes doivent être changées en fonction des déplacements de l'automobile. Bien que de tels appareils aient été installés dans des véhicules d'urgence, les entreprises, où devrait avoir lieu l'étape suivante de l'application de ce système, ne l'utilisent pas encore couramment.

La dernière fois que ces systèmes ont été examinés par le Census Bureau, aucun d'entre eux n'était suffisamment précis, économique et sûr pour pouvoir être utilisé dans les applications typiques relatives aux enquêtes. Cependant, puisqu'il est certain que leur précision ainsi que leur facilité d'exploitation s'amélioreront avec le temps et qu'ils deviendront moins coûteux, ils demeurent des options fascinantes pour l'avenir. Par exemple, des représentants de Magnavox ont prédit au cours d'une démonstration faite au Census Bureau que, d'ici l'an 2000, des récepteurs portatifs pour le système de positionnement global NAVSTAR avec une erreur de moins de 5 mètres devraient être disponibles, pour moins de $500 U.S. dans les magasins locaux de matériel d'escalade. Il se peut, qu'à ce moment, des modèles installables dans une automobile avec affichage numérique de cartes puissent aussi être disponibles chez les fournisseurs d'appareils électroniques pour automobiles. Dans le futur, il pourrait aussi être souhaitable de fournir aux intervieweurs des téléphones pour leur automobile et (ou) des téléavertisseurs, si le coût de ces appareils diminue. Cependant, la pratique courante, dans le cas des enquêtes, de demander aux intervieweurs de conduire leur propre automobile demeura un obstacle à l'utilisation de tout appareil qui doit être installé dans une automobile.

2.3 Interview sur place assistée par ordinateur

C'est la méthode de l'interview sur place assistée par ordinateur (IPAO) qui, de toutes les technologies de la phase 2 utilisées jusqu'ici pour les enquêtes-ménages, promet le plus. Cela fait plus de dix ans que le concept d'interviewer les répondants dans leur propre domicile ou dans leur propre bureau à l'aide d'ordinateurs portatifs a été reconnu, mais l'application pratique de cette méthode a été limitée jusqu'à ce que des ordinateurs personnels évolués à main ou portatifs deviennent d'un usage courant.

Bien que de petits essais d'IPAO aient été effectués auparavant [5, 8], ce n'est qu'en 1987 que l'on a constaté, pour la première fois, que cette technologie constituait une méthode manifestement viable d'effectuer la collecte de données auprès des ménages, à l'échelle d'un pays. En Europe, le Bureau central de la Statistique des Pays-Bas a commencé à utiliser de façon opérationnelle des ordinateurs à main en 1987 pour son enquête sur la population active [28, 58]. Aux États-Unis, la société National Analysts a commencé à effectuer la collecte de données, à l'échelle du pays, à l'aide de la méthode de l'IPAO pour l'enquête nationale sur l'alimentation de 1987 parrainée par le USDA [39]. Des essais importants de l'IPAO ont été effectués par le U.S. Census Bureau [31, 38], par le Office of Population Censuses and Surveys du Royaume-Uni ainsi que par le Research Triangle Institute [44].

En collaboration avec le National Center for Health Statistics, le Census Bureau a commencé, en 1988, à utiliser de façon continue l'IPAO pour l'enquête nationale sur la santé (ENS). Les 200 intervieweurs travaillant à l'ENS ont tous été pourvus d'un ordinateur portatif Gridlite Plus et on les a formés à leur utilisation. La première application est limitée à un supplément à l'ENS portant sur le SIDA et durant de 15 à 20 minutes. Les suppléments additionnels à l'ENS seront transférés de la méthode manuelle à l'IPAO à compter de 1990. On prévoit convertir les questions de base de l'ENS à l'IPAO dans le futur. Le Census Bureau compte aussi effectuer, en 1990, une importante étude de faisabilité portant sur l'IPAO lors de l'enquête sur la population active [55]. Des plans ont été élaborés, mais ils n'ont pas encore été approuvés et les fonds nécessaires n'ont pas été octroyés, pour fournir à tous les intervieweurs de l'EPA et de l'ENC le matériel nécessaire pour utiliser la méthode de l'IPAO, au milieu des années 90 [56].

Le développement de l'interview sur place assistée par ordinateur (IPAO) diffère de celui de l'interview téléphonique assistée par ordinateur (ITAO) sous au moins deux aspects importants.

Premièrement, la méthode de l'ITAO a été élaborée à l'origine dans le secteur privé pour des études de marché et dix années se sont écoulées avant que les organismes gouvernementaux effectuent leurs premières enquêtes à l'aide de l'ITAO [32]. Par contraste, les organismes gouvernementaux sont à l'avant-garde de l'élaboration de la méthode de l'IPAO à la fois comme parrains des essais d'IPAO effectués par des universités et des organismes du secteur privé et par l'élaboration de systèmes d'IPAO à leur propre usage. Cette situation peut être expliquée par l'importance de l'interview sur place pour les principales enquêtes effectuées par les organismes fédéraux ainsi que par les investissements relativement importants nécessaires pour entreprendre la collecte des données par IPAO.

Deuxièmement, la mise en application de l'IPAO dans les enquêtes gouvernementales s'effectue beaucoup plus rapidement que cela n'a été le cas pour l'ITAO. Cette dernière méthode était déjà relativement mûre et bien éprouvée avant que des organismes du gouvernement américain n'en considèrent sérieusement l'utilisation. Même alors, des organismes tels que le Census Bureau, Statistique Canada et ce qui était alors le Statistical Reporting Service du USDA (qui est maintenant le NASS) ont évalué avec soin ses effets sur la qualité des données, sur les estimations d'enquête ainsi que sur les coûts avant de l'utiliser pleinement [17, 32]. Par contraste, la mise en application de l'IPAO pour les enquêtes gouvernementales s'effectue rapidement sans recherches préalables pour en évaluer l'effet. Des essais de faisabilité ont démontré que presque tous les répondants jugent la méthode de collecte des données par IPAO acceptable et que la majorité des intervieweurs qui ont déjà de l'expérience dans le travail sur le terrain peuvent apprendre à l'utiliser [5, 28, 31, 38, 39, 44, 58], mais aucune évaluation complète des effets de l'IPAO sur la qualité des données et sur les estimations n'a été publiée jusqu'ici.

Un certain nombre de questions fondamentales d'ordre technologique relatives à l'IPAO demeurent encore sans réponse. Tout d'abord, "Quelles sont les meilleures méthodes pour transmettre les tâches et les questionnaires aux intervieweurs chargés de l'IPAO et pour retourner les données relatives aux interviews terminées à l'administration centrale?" Dans le cas de l'enquête nationale sur l'alimentation, la société National Analysts a fait appel à la poste, à la société UPS et à des services de messageries accélérés pour le transport des tâches et des données relatives aux interviews terminées [39]. Dans le cas de ces dernières données, elles ont été retournées sur disquettes. La société National Analysts avait pensé utiliser des services de télécommunications; dans ce cas, les intervieweurs auraient relié leur micro-ordinateur à leur téléphone à domicile pendant la nuit, mais cette méthode a été rejetée parce qu'elle présentait trop de risques, particulièrement dans les régions où les

services de la société Bell ne sont pas disponibles. Il ne fait aucun doute que pour la première utilisation de l'IPAO à l'échelle nationale, l'envoi des disquettes par la poste représentait un choix judicieux.

Par contraste, le Research Triangle Institute a utilisé les télécommunications automatisées avec succès lors de deux essais locaux d'IPAO réalisés dans le cadre de l'enquête utilisant la méthode d'évaluation de l'exposition totale (MEET ("TEAM")) menée pour la U.S. Environmental Protection Agency [44]. Pour l'enquête sur la population active des Pays-Bas, on a aussi utilisé avec succès des appareils de télécommunication afin de transmettre les données recueillies par IPAO [58]. La possibilité d'utiliser des moyens de télécommunication rapides entre les administrations centrales et les intervieweurs présente de nombreux avantages, particulièrement dans le cas d'enquêtes comportant des délais serrés et des périodes d'interview limitées, comme l'enquête sur la population active aux États-Unis et l'Enquête sur la population active au Canada. Le Census Bureau prévoit élaborer un système intégré d'ITAO-IPAO dans lequel les dossiers pourraient être échangés rapidement entre les intervieweurs utilisant la méthode de l'ITAO dans des bureaux centralisés et les intervieweurs utilisant la méthode de l'IPAO dispersés dans tout le pays. Dans le cas des intervieweurs utilisant l'IPAO, il reste à démontrer qu'un tel système est réalisable dans toutes les parties du pays. Le transfert des données par télécommunication pose aussi des problèmes additionnels (réels et apparents) relatifs à la protection des données confidentielles. Dans le cas de la première utilisation véritable de l'IPAO pour le supplément à l'ENS portant sur le SIDA, le Census Bureau utilise la méthode plus prudente qui consiste à expédier les disquettes par la poste.

La deuxième question à résoudre est la suivante: "Comment peut-on le mieux adapter le logiciel d'IPAO aux besoins spéciaux que comporte la réalisation d'interviews à domicile?" Le logiciel utilisé jusqu'à ce jour pour la majorité des applications d'IPAO a été conçu pour les interviews téléphoniques assistées par ordinateur ou pour les interviews sur place dans des bureaux. La présentation des questions, les modes d'introduction des données et les commandes à l'intention des intervieweurs, qui sont acceptables dans ces cadres plus rigides, peuvent ne pas être aussi appropriés lorsque l'interview a lieu au domicile du répondant ou "sur le pas de la porte", les conditions étant alors plus difficiles et moins favorables à la concentration. Bien que des lignes directrices générales relatives à la réalisation des questionnaires pour les interviews assistées par ordinateur aient été proposées [19, 20, 33], on n'a pas encore déterminé les fonctions particulières, du point de vue logiciel, qui seraient les plus utiles aux intervieweurs utilisant la méthode de l'IPAO, ou il n'y a pas encore d'accord véritable à ce sujet.

Une troisième question qui reste à régler dans le cas de l'IPAO et, peut-être, la principale à l'heure actuelle est la suivante: "De quel type de matériel les intervieweurs qui utilisent la méthode de l'IPAO devraient-ils disposer?" Bien que le Bureau central de la Statistique des Pays-Bas ait élaboré son système d'IPAO pour des micro-ordinateurs à main avec de petits écrans de visualisation, entre autres le Epson PX-4, pour les essais d'IPAO réalisés aux États-Unis, on a généralement utilisé des ordinateurs portatifs compatibles avec le matériel IBM tels que le Toshiba T-1100+ et le Gridlite Plus. Les ordinateurs portatifs de ce type sont munis de touches de fonction programmables, d'écran et de clavier de dimensions standard et habituellement ils pèsent au moins 10 livres. Les ordinateurs portatifs disponibles actuellement peuvent peser jusqu'à 16 livres. Ils ont une mémoire vive plus considérable, un disque rigide et une vitesse de traitement plus rapide. De plus, ils peuvent être nécessaires quand les questionnaires sont enregistrés dans deux langues (français-anglais ou anglais-espagnol), quand on désire effectuer des interviews pour différentes enquêtes lors du même déplacement et quand on veut tenir à jour des registres des appels.Ce poids est excessif pour les intervieweurs sur le terrain. Le coût de ces micro-ordinateurs est aussi prohibitif actuellement pour la majorité des applications.

Nous ne savons pas encore si la majorité des intervieweurs peuvent tolérer le fardeau que représente un micro-ordinateur portatif comme équipement permanent utilisé pour leur travail, même s'il ne pèse que dix livres. Les expériences dont les résultats ont été publiés et qui portent sur des tests récents d'IPAO sont contradictoires [31, 38, 44]. Il se peut que seule l'expérience nous permette de répondre à cette question, ou que le cas se règle quand des micro-ordinateurs portatifs plus légers et à haute performance seront disponibles.

Les micro-ordinateurs portatifs standard présentent aussi d'autres inconvénients dans le cas des interviews sur le terrain. Tout d'abord, ils sont mal protégés contre les chocs, les liquides renversés, la poussière ou l'humidité excessive et les températures extrêmes. Deuxièmement, comme il s'agit d'ordinateurs personnels universels, il se peut que des intervieweurs (ou des membres de leur famille) soient tentés de les utiliser à d'autres fins que pour remplir les tâches reliées aux interviews. Cela pourrait compromettre la confidentialité des données fournies par les répondants ou possiblement effacer des programmes ou des fichiers de données essentiels. Troisièmement, comme ces ordinateurs portatifs sont des ordinateurs universels, ils sont aussi des cibles tentantes pour les voleurs, particulièrement quand on les transporte dans des régions où le taux de criminalité est élevé. Finalement, puisque les ordinateurs portatifs doivent être en position horizontale et supportés par en-dessous quand on les utilise, ils ne sont pas appropriés dans le cas des interviews "sur le pas de la porte" ou "à travers le moustiquaire".

Dans leur commercialisation, on présente certains appareils portatifs de reconnaissance de caractères manuscrits (RCM) comme de meilleures solutions aux problèmes reliés à la collecte des données sur le terrain que ne le sont les ordinateurs portatifs. Le boîtier de ces appareils comprend une membrane en plastique, il peut supporter des conditions environnementales plus rigoureuses et une manipulation moins délicate. Le fait que ces appareils soient conçus pour remplir une tâche spéciale et leur apparence particulière devraient décourager à la fois le vol et leur emploi pour d'autres travaux que ceux reliés aux enquêtes. Comme leur périphérique, de type bloc-note, servant à l'introduction des données pèsent entre 3 et 4 livres, ils peuvent être tenus d'une seule main comme une plaquette à pince pendant que les inscriptions sont faites avec l'autre main. Les batteries et les autres éléments lourds peuvent être transportés dans un sac à bandoulière et reliés, par un câble, au bloc-note. Comme ces appareils émulent l'introduction de données à l'aide de questionnaires sur papier, il paraît que la formation des intervieweurs est simple et que l'emploi de ces appareils ne cause pas de préoccupations. Ces appareils peuvent aussi remplir diverses fonctions spécialisées comme l'enregistrement de totaux, de cartes, de diagrammes et de signatures.

Pour introduire les données, on utilise un stylo ou un stylet pour faire une marque dans une case ou pour écrire des caractères dans des espaces déterminés. Ces inscriptions sont ensuite converties en caractères ASCII par le système. Afin de confirmer l'introduction des données, ces caractères sont ensuite affichés en caractères informatiques standard. Plus tard, on peut utiliser les mêmes appareils pour transmettre les données relatives à des interviews terminées à un ordinateur dans un bureau central, à partir d'une ligne téléphonique.

Deux systèmes dont la démonstration a été faite au Census Bureau utilisent des méthodes fort différentes pour l'introduction et la reconnaissance des caractères. Avec le **Scriptwriter** de DATAKECH, l'intervieweur utilise un stylo à bille ordinaire pour remplir une formule en papier placée sur une surface de plastique sensible à la pression. La formule en papier remplie devient une copie de sécurité imprimée des données enregistrées dans l'ordinateur. On ne peut introduire que des chiffres, des caractères de ponctuation ordinaires et des majuscules d'imprimerie avec un seul caractère par espace et ces caractères doivent être formés d'une façon prescrite.

L'appareil **Write-Top** de LINUS utilise un stylet spécial pour écrire sur une plaque de verre sensibilisé qui se trouve au-dessus de l'affichage à cristaux liquides du questionnaire ou de la formule. Quand le stylet touche la plaque de verre, cette dernière fonce ce qui donne visuellement l'impression d'écrire. Les caractères écrits paraissent pour une courte période puis sont remplacés, à peu près au même endroit, par le caractère informatique standard équivalent, comme l'ordinateur l'a reconnu. L'introduction et la reconnaissance des caractères dépendent de l'utilisateur. Il est permis d'utiliser les caractères majuscules et minuscules et l'intervieweur a beaucoup de latitude pour ce qui est de la façon dont il forme les caractères. Cependant, l'intervieweur doit "apprendre" à l'appareil **Write-Top** qui lui a été fourni comment reconnaître la façon dont il forme chaque lettre, en majuscule et en minuscule. Cette opération prend environ 2 à 2.5 heures.

Il se peut que cette technologie relativement nouvelle puisse avoir diverses utilisations spéciales pour la collecte des données relatives aux enquêtes et aux recensements, y compris l'inscription des îlots, l'introduction de données sur le terrain à partir d'archives imprimées et la surveillance des interviews téléphoniques. Cependant, au stade actuel de leur mise au point, il ne semble pas que ces appareils portatifs de RCM puissent remplacer de façon acceptable les ordinateurs portatifs utilisés pour l'IPAO dans le cas des principales enquêtes-ménages. Leurs écrans de visualisation sont actuellement trop petits ou trop difficiles à lire; ils n'ont pas été conçus pour traiter l'établissement de la liste des membres du ménage ou les formules complexes de plusieurs pages et ils ne fournissent pas les fonctions dont on dispose habituellement quand on emploie des ordinateurs, comme l'enchaînement automatique des questions, le contrôle des gammes de valeurs permises et de la cohérence, des écrans de consultation intégrés ainsi que des questions libellées en fonction des réponses antérieures. En résumé, il s'agit surtout d'appareils de saisie des données qui conviennent bien au traitement de formules simples de collecte de données.

Si ce sont leurs propriétés ergonomiques qui constituent les avantages principaux des appareils de RCM, nous pouvons nous demander s'il ne serait pas possible de construire du matériel d'IPAO qui posséderait l'affichage pleine page et la puissance de traitement d'un ordinateur portatif mais qui serait aussi durable, portatif et adaptable aux interviews "sur le pas de la porte" que les appareils de RCM. Le Census Bureau étudie cette option avec des entreprises qui se spécialisent dans la conception ou la fabrication d'ordinateurs portatifs spécialisés. Par exemple, bien qu'il semble qu'un écran pleine page soit essentiel pour les enquêtes-ménages, il se peut qu'un clavier standard soit inutile ou qu'il constitue même un désavantage. Bryant [36] a proposé de modifier les claviers afin que les intervieweurs puissent les utiliser facilement d'une seule main, car peu d'intervieweurs connaissent la dactylographie au toucher.

Il se peut que les organismes qui effectuent des enquêtes trouvent qu'il n'existe pas une meilleure solution en matière de matériel informatique pour l'IPAO. Il se peut que les ordinateurs portatifs de dimension standard soient appropriés dans le cas des interviews de longue durée où les intervieweurs sont habituellement assis dans le domicile des répondants. De plus petits modèles avec des claviers disposés de façon à pouvoir être utilisés d'une seule main pourraient être plus appropriés dans le cas d'interviews de courte durée menées "sur le pas de la porte". Les appareils portatifs de reconnaissance de caractères manuscrits peuvent être le meilleur choix pour les formules d'une seule page, pour l'établissement des listes d'îlot et pour d'autres applications. Plutôt que de chercher l'outil d'IPAO général ultime, il se peut que l'on doive constituer un arsenal composé de divers matériels informatiques, parmi lesquels on choisirait l'appareil le plus approprié à chaque application.

Quelles que soient les solutions qui seront trouvées aux besoins imposés par l'utilisation de la méthode de l'IPAO, tant du point de vue matériel que du point de vue logiciel, il

semble probable que cette nouvelle technologie aura un effet considérable sur les opérations sur le terrain dans le cas des enquêtes-ménages. Par exemple:

1. Les surveillants des intervieweurs sur le terrain et les chefs d'équipe devront être choisis non seulement pour leur entregent mais aussi pour leur capacité à régler les problèmes techniques relatifs au matériel ainsi qu'au logiciel d'IPAO et aux télécommunications.

2. La formation des intervieweurs, tant la formation initiale que celle qui porte sur une enquête particulière, devra comprendre de nouveaux sujets techniques tels que l'utilisation et la maintenance des micro-ordinateurs utilisés pour l'IPAO, l'interview à l'aide de la méthode de l'IPAO et la transmission des tâches et des données. En même temps, des progiciels d'enseignement assistés par ordinateur portant sur ces sujets et sur d'autres pourront être distribués sur disquettes ou transférés au moyen des ordinateurs des administrations centrales pour la formation à domicile des intervieweurs.

3. La surveillance des intervieweurs changera tant pour ce qui est de la façon dont elle est effectuée que de ce qui est surveillé. Les contrôles par lots et (ou) le survol, par ordinateur, des résultats et des fichiers des registres des visites remplaceront le contrôle visuel des questionnaires remplis et des registres des visites. L'enchaînement automatique des questions et les contrôles en direct pourront éliminer de nombreux types bien connus d'erreurs dues aux intervieweurs, mais il se peut que les surveillants aient à élaborer de nouvelles méthodes pour reconnaître les intervieweurs qui ont besoin d'un complément de formation en rapport avec les commandes et les opérations de l'IPAO.

4. Le contrôle à l'arrivée ainsi que l'examen initial des données recueillies par IPAO et transmises par télécommunication seront effectués surtout par des ordinateurs. Il restera très peu de travail pour les employés de bureau oeuvrant aux enquêtes sur le terrain soit dans les administrations centrales, soit dans les régions; tandis que la demande pour du personnel de soutien technique croîtra.

5. Les rapports qui existent entre les administrations centrales et les bureaux régionaux, d'État ou provinciaux (quand il y en a) pourront changer avec les nouvelles options en matière de communication et de transmission. Par exemple, les données relatives aux interviews terminées pourraient être expédiées directement aux administrations centrales plutôt que par l'intermédiaire de bureaux locaux et il se pourrait que les administrations centrales puissent envoyer directement des messages aux intervieweurs.

En résumé, il se peut que l'IPAO introduise de nombreux changements dans la façon dont les interviews sur place sont affectées, menées, surveillées, contrôlées à l'arrivée, examinées et administrées et ces changements auront un effet sur le genre d'employé requis et sur la façon dont ce personnel peut être employé et dirigé le plus efficacement possible.

2.4 Technologie vocale

L'arrivée de nouveaux groupes de conseillers pour les organismes s'occupant d'enquêtes constitue un indice de l'importance croissante de la technologie de pointe pour la collecte de données statistiques. Le National Bureau of Standards (NBS) a récemment préparé, pour le Census Bureau, un rapport intitulé "Automation of Data Capture for the Census in the Year 2000" [36]. Si les prévisions contenues dans ce rapport se réalisent, l'ITAO et l'IPAO comme nous les connaissons actuellement ne constituent que des modes transitoires de collecte des données.

Dans son rapport, le NBS affirme que le mode le plus naturel de collecte de données pour les humains est ni le papier, ni les claviers mais bien la parole. Il recommande que le Census Bureau étudie la possibilité d'ajouter des techniques de pointe en matière de technologie vocale aux améliorations qu'il apportera à la méthode de l'ITAO. La technologie vocale comprend à la fois la simulation et la reconnaissance de la parole. Des applications plus perfectionnées de cette dernière technologie se prolongent jusqu'à l'analyse de dialogues et l'intelligence artificielle [43].

Le NBS a proposé deux objectifs pour l'application de la technologie vocale à la collecte de données. Le premier est de "réaliser des interviews téléphoniques entièrement automatisées" (traduction libre), c'est-à-dire des interviews téléphoniques menées sans intervieweur humain. Le second objectif est d'élaborer la technologie vocale comme "mode [auxiliaire] de commande (ou) de contrôle pour les systèmes portatifs de saisie de données, ce qui n'obligerait pas les intervieweurs à maîtriser aussi bien l'utilisation d'un clavier." (Traduction libre.)

Où ces options se trouvent-elles dans le futur? Plus on connaît le travail sur le terrain en matière d'enquêtes et moins on connaît la technologie vocale, plus il pourrait sembler que ces options sont éloignées (ou improbables). Le NBS les a proposées comme solutions de rechange réalistes pour le recensement de l'an 2000, mais certains exemples d'application de cette technologie existent déjà. Les plus prometteurs sont ceux qui peuvent être utilisés pour les enquêtes auprès des établissements. Le Bureau of Labor Statistics, comme nous l'avons déjà indiqué, fait déjà l'essai d'une forme de base de la technologie vocale, le traitement des transactions, pour la partie du programme des statistiques courantes sur l'emploi dans laquelle on utilise un téléphone à clavier pour introduire les données. Le BLS prévoit aussi faire, d'ici peu, l'essai de la reconnaissance vocale pour remplacer la saisie des données à l'aide d'un téléphone à clavier.

L'utilisation de la technologie vocale pour les enquêtes-ménages présente un défi plus considérable, bien qu'on puisse déjà trouver des exemples d'utilisation de cette technologie dans le secteur privé. Feinberg [12] a décrit des essais de numérisation des signaux vocaux pour enregistrer les réponses à des questions "ouvertes" (et les questions elles-mêmes si on le désire) lors d'interviews réalisées par ITAO pour des études de marché. Dans ce cas, il faut, bien entendu, obtenir l'approbation préalable du répondant. Les réponses numérisées peuvent faire l'objet d'un classement recoupé selon les caractéristiques des répondants et on peut les faire entendre au parrain de l'enquête. Dans le futur, on pourra les soumettre à une analyse visant à reconnaître la parole. Stonecash [51] a présenté un des nombreux systèmes de sondage électoral en usage actuellement qui emploient une méthode d'interview téléphonique entièrement automatisée (sans intervieweur) avec composition automatique, enregistrements automatiques et réponse par téléphone à clavier. Les faibles coûts d'exploitation de ces systèmes entièrement automatisés favoriseront sans aucun doute leur utilisation et leur développement, mais ils n'ont pas encore atteint des taux de réponse qui s'approchent de la norme pour la majorité des enquêtes gouvernementales. Le facteur clé dans l'utilisation de la technologie vocale pour les enquêtes-ménages semblerait être l'acceptation par les répondants. Jusqu'à ce qu'il y ait d'autres progrès relatifs à cette nouvelle technologie et qu'elle soit l'objet de plus d'essais sur le terrain, il est difficile de prédire son rôle ultime dans la collecte des données auprès des ménages et il ne faut pas en préjuger.

3. Résumé et conclusions

D'après les expériences réalisées jusqu'à ce jour, l'effet des technologies de pointe actuelles et à venir sur la collecte des données d'enquête se poursuivra apparemment pendant encore des décennies, peut-être à un rythme accéléré. Bien que certains

organismes qui effectuent des enquêtes s'efforcent encore de mettre en application la méthode de l'ITAO et de comprendre toutes ses conséquences relatives aux estimations d'enquête, ils doivent se préparer à une mise en application beaucoup plus rapide de la méthode de l'IPAO. Et alors qu'ils sont à peine à l'aise suite au remplacement du papier par des modes de collecte des données qui font appel à un clavier, des conseillers bien informés leur suggèrent de passer à l'étape suivante qui est celle de la technologie vocale. D'autres options, telles que les systèmes de radionavigation, les questionnaires informatisés à remplir soi-même et les appareils portatifs de reconnaissance de caractères manuscrits, causent encore plus de confusion dans un avenir qui offre de nombreux choix possibles.

Si cette évaluation est exacte, les organismes gouvernementaux qui effectuent la collecte de données doivent se préparer à des changements technologiques continus (ou du moins fréquents) dans leurs travaux de collecte des données; ces changements ressembleront aux améliorations périodiques mieux connues qu'ils doivent apporter à leurs systèmes informatiques. Le fait de changer les technologies utilisées pour la collecte des données pourrait se révéler plus difficile parce que l'emploi de ces technologies soulève très fréquemment des questions d'ordre organisationnel et méthodologique. Cela est très manifeste dans le cas de la mise en application de l'ITAO par les gouvernements lors de la phase 1 et cela pourrait l'être moins dans le cas de l'application de l'IPAO et de la technologie vocale dans la phase 2. Plutôt que de considérer les technologies de collecte des données comme des technologies statiques (ou du moins des technologies qui changent rarement), il pourrait se révéler nécessaire de les considérer comme des technologies dynamiques. Dans un monde où les changements sont fréquents, il se peut que les organismes gouvernementaux trouvent nécessaire d'élaborer des méthodes plus efficientes et plus opportunes pour reconnaître, évaluer, comparer et mettre en application de nouvelles technologies de collecte des données à mesure que ces technologies évoluent.

BIBLIOGRAPHIE

[1] Bailar, Barbara A., "The Effects of Rotation Group Bias on Estimates from Panel Surveys," Journal of the American Statistical Association, Vol. 70, No. 349, mars 1975, pp. 23-30.

[2] Bass, Robert T. et Robert D. Tortora, "A Comparison of Centralized CATI Facilities for An Agricultural Labor Survey," dans Robert M. Groves et coll. (éd.), Telephone Survey Methodology, New York, Wiley Press, en voie de rédaction.

[3] Biemer, Paul, David W. Chapman et Charles Alexander, "Some Research Issues in Random-Digit Dialing Sampling and Estimation," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 71-86.

[4] Berry, Sandra H. et Diane O'Rourke, "Administrative Designs for Centralized Telephone Survey Centers: Implications of the Transfer to CATI," dans Robert M. Groves et coll. (éd.) Telephone Survey Methodology, New York, Wiley Press, pp. 457-474.

[5] Birkett, N.J., "Computer-Aided Personal Interviewing: A New Technique for Data Collection in Epidemiologic Surveys," American Journal of Epidemiology, Vol. 127, No. 3, 1988, pp. 684-690.

[6] Bushery, John M. et Lawrence Cahoon, "Labor Force Characteristics," dans U.S. Census Bureau, Evaluation of CATI Data Quality and Costs in the Current Population Survey, CATI Research Report No. CPS-2 of the Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, septembre 1988, pp. 41-49.

[7] Catlin, Gary, Susan Ingram et Lecily A. Hunter, "Les effets de l'ITAO sur les coûts et la qualité des données," dans le présent ouvrage.

[8] Danielsson, L., et P.A. Maarstad, "Statistical Data Collection with Handheld Computers - A Test in Computer Price Index." Rapport non publié du Bureau de la statistique de la Suède, Orebro, Suède, 1982.

[9] Dillman, Don A. et John Tarnai, "Administrative Issues in Mixed Mode Surveys," dans Robert M. Groves et coll. (éd.), Telephone Survey Methodology, New York, Wiley Press, pp. 509-528.

[10] Drew, J. Douglas, G. Hussain Choudhry et Lecily A. Hunter, "Nonresponse Issues in Government Telephone Surveys," dans Robert M. Groves et coll. (éd.) Telephone Survey Methodology, New York, Wiley Press, pp. 233-246.

[11] Dutka, Solomon et Lester Frankel, "Sequential Survey Design Through the Use of Computer Assisted Telephone Interviewing," Proceedings of the Section on Business and Economic Statistics, American Statistical Association, 1980, pp. 73-76.

[12] Feinberg, Barry, "A&S Voice CATI." Communication présentée lors de la Conférence de l'American Association for Public Opinion Research, mai 1988.

[13] Fink, James C., "CATI's First Decade: The Chilton Experience," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 153-168.

[14] Groves, Robert M. et Robert L. Kahn, Surveys by Telephone, New York, Academic Press, 1979.

[15] Groves, Robert M. et James M. Lepkowski, "Dual Frame, Mixed Mode Survey Designs," Journal of Official Statistics, Vol. 1, No. 3, 1985, pp. 263-286.

[16] Groves, Robert M. et James M. Lepkowski, "Cost and Error Modeling for Large-Scale Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 330-357.

[17] Groves, Robert M. et William L. Nicholls II, "The Status of Computer-Assisted Telephone Interviewing: Part II -- Data Quality Issues," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 117-134.

[18] Harding, Dan, Maurice J. Moroney et Albert D. Frost, "Radionavigation Systems Analysis Evaluation Program for the 1990 Census." Communication présentée lors de la réunion technique nationale de l'Institute of Navigation, janvier 1984.

[19] House, Carol C., "Questionnaire Design with Computer-Assisted Telephone Interviewing," Journal of Official Statistics, Vol. 1, No. 2, 1985, pp. 209-219.

[20] House, Carol C. et William L. Nicholls II, "Questionnaire Design for CATI: Design Objectives and Methods," dans Robert M. Groves et coll. (éd.), Telephone Survey Methodology, New York, Wiley Press, pp. 421-436.

[21] Hubble, David L. et Bruce E. Wilder, "Preliminary Results from the National Crime Survey (NCS) CATI Experiment," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988, en voie de rédaction.

[22] Kefron, J.H. et R.O. Nelson, "Centralized Long-Distance Interviewing -- A Revolution in Data Collection." Communication présentée lors du 12e congrès annuel de l'Advertising Research Foundation, octobre 1966.

[23] Kiesler, Sara et Less S. Sproull, "Response Effects in Electronic Surveys," Public Opinion Quarterly, Vol. 50, No. 3, automne 1986, pp. 402-413.

[24] Lyberg, Lars, "Administration of Telephone Surveys" dans Robert M. Groves et coll. (éd.), Telephone Survey Methodology, New York, Wiley Press, pp. 453-456.

[25] Martel, Col. John, "Status of Radionavigation Systems," Proceedings of the Surface Transportation Users Conference on Navigation, Washington, D.C., novembre 1983.

[26] Marquis, Kent et Richard Blass, "Nonsampling Error Considerations in the Design and Operation of Telephone Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 301-329.

[27] Massey, James T., Kent H. Marquis et Robert D. Tortora, "Methodological Issues Related to Telephone Surveys by Federal Agencies," Proceedings of the Social Statistics Section, American Statistical Association, 1982, pp. 63-72.

[28] Bureau central de la Statistique des Pays-Bas, Automation in Survey Processing, Select Report 4, Voorburg, Pays-Bas, Bureau central de la Statistique, 1987.

[29] Nelson, R.O., B.L. Peyton et B.Z. Bortner, "Use of an On-Line Interactive System: Its Effects on the Speed, Accuracy, and Costs of Survey Results." Communication présentée lors du 18e congrès de l'Advertising Research Foundation, New York, 1972.

[30] Nicholls II, William L., "Experiences with CATI in a Large-Scale Survey," Proceedings of the Survey Research Methods Section, American Statistical Association, 1978, pp. 9-17.

[31] Nicholls II, William L., "Discussion" des communications présentées par Rothschild et Wilson et par Sebestik et coll., Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 340-342.

[32] Nicholls II, W.L. et Groves, R.M., "The Status of Computer-Assisted Telephone Interviewing: Part I -- Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 93-115.

[33] Nicholls II, William L. et Carol C. House, "Designing Questionnaires for Computer-Assisted Interviewing," Proceedings of the Bureau of the Census Third Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1987, pp. 95-111.

[34] Nicholls II, William L., Michael E. Haas et Richard F. Blass, "Cost Models for Computer-Assisted Interviewing in the CPS," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988.

[35] Pafford, Bradley V. et Dick Coulter, "Use of Historical Data in a Current Interview Situation," Proceedings of the Bureau of the Census Third Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1987, pp. 281-298.

[36] Pallet, D.S. (éd.), Automation of Data Capture for the Census in the Year 2000. Rapport final présenté au U.S. Bureau of the Census par le National Bureau of Standards en vertu de l'accord inter-organismes pour l'exercice financier 1987.

[37] Remington, Patrick L., Meredith Y. Smith, David F. Williamson, Robert F. Anda, Eileen M. Gentry et Gary C. Hogelin, "Design, Characteristics, and Usefulness of the State-Based Behavioral Risk Factor Surveillance: 1981-87," Public Health Reports, Vol. 103, No. 4, juillet-août 1988, pp. 366-375.

[38] Rice Jr., Stewart C., Robert A. Wright et Ben Rowe, "Development of Computer Assisted Personal Interview For the National Health Interview Survey," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988, en voie de rédaction.

[39] Rothschild, Beth B. et Lucy B. Wilson, "Nationwide Food Consumption Survey 1987: A Landmark Personal Interview Survey Using Laptop Computers," Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 347-356.

[40] Rustemeyer, Anitra et Arnold Levin, "Report on a Telephone Survey Using Computer Assistance," Rapport non publié du U.S. Department of Commerce, Bureau of the Census, 1977.

[41] Rustemeyer, Anitra, Gerald H. Shure, Miles S. Rogers et Robert J. Meeker, "Computer-Assisted Telephone Interviewing: Design Considerations," Proceedings of the Survey Research Methods Section, American Statistical Association, 1978, pp. 1-6.

[42] Schwanz, Dennis, Ed Montfort et John Cannon, "Analysis of Operational Issues: 1987 AHS CATI," CATI Research Report No. AHS-1, U.S. Bureau of the Census, Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, août 1988.

[43] Science, "A Landmark in Speech Recognition", Vol. 240, 17 juin 1988, p. 1615.

[44] Sebestik, Jutta, Harvey Zelon, Dale DeWitt, James M. O'Reilly et Kevin McGowan, "Initial Experiences with CAPI," Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 357-365.

[45] Sebold, Janice, "Survey Period/Length, Unanswered Numbers, and Nonresponse in Telephone Surveys," dans Robert M. Groves et coll. (éd.) Telephone Survey Methodology, New York, Wiley Press, pp. 247-256.

[46] Shanks, J. Merrill et Robert Tortora, "Beyond CATI: Generalized and Distributed Systems for Computer-Assisted Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 358-371.

[47] Shuldiner, Herbert, "Computerized Navigator for Your Car," Popular Science, juin 1985, pp. 64-67.

[48] Spaeth, Mary A., "Telephone Interviewing Facilities at Survey Research Organizations," Survey Research, Vol. 11, Nos. 3-4, été-automne 1979, pp. 21-25.

[49] Spaeth, Mary A., "CATI Facilities at Survey Research Organizations," Survey Research, Vol. 18, Nos. 3-4, été-automne 1987, pp. 18-22.

[50] Stock, J. Steven, "How To Improve Samples Based on Telephone Listings," Journal of Advertising Research, Vol. 2, No. 3 (septembre 1962), pp. 555-564.

[51] Stonecash, Jeffrey, "Automated Dialing, Automated Recordings, and the Effects on Samples." Communication présentée lors de la conférence de l'American Association for Public Opinion Research, mai 1988.

[52] Swann, T.C., "Electronic Data Collection in the Petroleum Supply Reporting System." Communication présentée lors de la rencontre du Committee on Energy Statistics de l'American Statistical Association, 28-29 avril 1988.

[53] Thornberry, Owen T. et Gail S. Poe, "NCHS Research on the Telephone Interview: Some Observations," Proceedings of the Social Statistics Section, American Statistical Association, 1982, pp. 296-301.

[54] Tortora, Robert D., "CATI in an Agricultural Statistical Agency," Journal of Official Statistics, Vol. 1, No. 2, 1985, pp. 301-314.

[55] U.S. Census Bureau, Computer Assisted Interviewing: Research, Development, and Implementation. Rapport du Central Planning Committee for Computer Assisted Interviewing, 1987.

[56] U.S. Census Bureau, Final Report on Planning for a Post-1990 Redesign of the Current Population Survey. Rapport du Oversight Committee of the Joint BLS-Census Committee on the Current Population Survey, avril 1988.

[57] U.S. Census Bureau, Evaluation of CATI Data Quality and Costs in the Current Population Survey, CATI Research Report No. CPS-2 of the Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, septembre 1988.

[58] van Bastelaer, Alois, Frans Kessemakers et Dirk Sikkel, "Data Collection with Hand-Held Computers: Contributions to Questionnaire Design," Journal of Official Statistics, Vol. 4, No. 2, 1988, pp. 141-154.

[59] Waksberg, Joseph, "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Vol. 73, No. 361, mars 1976, pp. 40-46.

[60] Werking, George, Alan R. Tupek et Richard Clayton, "CATI and Touchtone Self-Response for Establishment Surveys," Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1988, pp. 273-290.

[61] Werking, George, Richard Clayton, Richard Rosen et Debbie Winter, "Conversion from Mail to CATI in the Current Employment Statistics Program," Proceedings of the Survey Research Methods Section, American Statistical Association, 1988, en voie de rédaction.

LES EFFETS DE L'ITAO SUR LES COÛTS ET LA QUALITÉ DES DONNÉES

G. CATLIN, S. INGRAM ET L. HUNTER,

Statistique Canada

1.0 INTRODUCTION

L'utilisation accrue de l'ITAO dans les organisations d'enquête au cours des dernières années est en partie attribuable au fait que cette méthode devait améliorer la qualité des données d'enquête par téléphone, réduire le coût de la collecte des données et raccourcir la durée des interviews. Les premières publications concernant l'ITAO semblait promettre tous ces avantages mais peu de témoignages écrits ont permis d'indiquer que l'ITAO a répondu aux attentes.

Les études comparant la méthode ITAO et la méthode faisant appel aux questionnaires sur papier ont été publiées par Coulter (1985), Groves et Mathiowetz (1984), Harlow, Rosenthal et Zeigler (1985) et House (1984). Dans une série de deux articles analysant la recherche sur les effets de l'ITAO, Nicholls et Groves (1986) ont résumé d'autres comparaisons non publiées ainsi que des indications subjectives. La plupart des données présentées s'appuient sur des enquêtes effectuées avec des échantillons restreints, sur des études ayant un plan expérimental mal contrôlé ou sur les premières tentatives d'application de la méthode ITAO par des organisations non familiarisées avec celle-ci.

Cette étude visait à évaluer les effets de l'ITAO sur les coûts d'enquête et la qualité des données dans un milieu d'interviews centralisé, dans des conditions relativement contrôlées et avec une étude suffisamment longue et importante pour permettre des résultats plus généralisables.

Les mêmes enquêteurs ont effectué la même enquête à partir d'un emplacement commun en utilisant la méthode ITAO une semaine et la méthode du questionnaire sur papier l'autre semaine. Le test se déroulait sur une vaste échelle. Entre février et septembre 1987, on a effectué des interviews auprès de plus de 10 000 ménages, avec à peu près le même nombre d'interviews dans les deux méthodes.

Ce test était effectué dans le cadre d'un projet de recherche conjoint de Statistique Canada et du U.S. Bureau of the Census (USBC). Statistique Canada a utilisé le logiciel de recensement du système ITAO du USBC et évalué l'effet de l'ITAO en se servant de l'enquête sur la population active (EPA) comme véhicule. L'EPA est similaire au Current Population Survey effectué par le USBC avec des questions portant sur l'emploi et les activités de recherche d'emploi se rapportant à une période de référence d'une semaine.

Les résultats initiaux de cette enquête sur les effets de l'ITAO sur la qualité des données et les coûts d'enquête pendant la période de collecte des données ont déjà fait l'objet de rapports (Catlin et Ingram, 1988; Catlin, Ingram et Hunter, 1988). L'information présentée compare les taux de réponses, les mesures de qualité des interviews, les estimations d'enquêtes, les coûts, les réactions des enquêteurs et l'évaluation du logiciel. Avant d'examiner ces résultats, nous allons d'abord passer en revue la méthodologie utilisée pour cette étude.

2.0 MÉTHODOLOGIE

L'étude s'est déroulée en Ontario. La population cible était comparable à celle de l'enquête sur la population active à savoir la population civile hors institution. Les ménages sans téléphone ainsi que les ménages collectifs de 10 personnes ou plus étaient également exclus. En Ontario, les ménages sans téléphone représentent approximativement deux pour cent des ménages et les ménages collectifs, moins de un pour cent des ménages (Statistique Canada, 1984).

Pour les deux méthodes, on a sélectionné l'échantillon à l'aide de la composition téléphonique aléatoire. On a choisi un échantillon systématique des banques actives d'après l'information obtenue de Bell Canada et d'autres sources, et un numéro à deux chiffres généré au hasard a été ajouté afin de former un numéro de téléphone complet. Un mois avant la collecte des données, on a effectué un examen sélectif de la composition téléphonique aléatoire pour identifier les ménages à l'aide d'un instrument de l'ITAO. Les ménages étaient affectés au hasard à l'interview selon la méthode ITAO ou selon la méthode du questionnaire sur papier et la même méthode d'interview était appliquée pour toute la durée de l'enquête.

Un échantillon d'environ 1,000 ménages par mois pour chaque méthode d'interview était visé. En février et en mars, on introduisait environ 333 ménages (2 groupes de rotation) pour chaque méthode. En avril et en mai, on ajoutait 167 ménages (1 groupe de rotation), ce qui a amené chaquea amené chaque échantillon au nombre requis de 1 000. De juin à septembre, un sixième de l'échantillon était remplacé chaque mois comme le veut l'usage dans l'EPA. Pendant la dernière semaine de septembre, on a effectué une nouvelle interview auprès du quart de l'échantillon de septembre, et ce, pour les deux méthodes.

L'interview a été effectuée à partir d'installations téléphoniques centralisées au bureau principal de Statistique Canada à Ottawa. Tous les enquêteurs utilisaient les deux méthodes afin de réduire l'influence de la variabilité de l'enquêteur sur les résultats du test. Les enquêteurs recevaient une formation d'une semaine sur les concepts de la population active et sur les procédures relatives à la méthode faisant appel aux questionnaires sur papier et à la méthode ITAO. Chaque mois, on employait deux superviseurs et de 8 à 11 enquêteurs.

Deux périodes d'interviews d'une semaine chacune étaient requises chaque mois, une pour les interviews selon la méthode ITAO et l'autre pour la méthode du questionnaire sur papier. Chaque mois, les interviews se déroulaient pendant les mêmes deux semaines; toutefois, les méthodes d'interviews assignées à chaque semaine alternaient d'un mois à l'autre. Les interviews se faisaient entre 8 h 30 et 21 h 30 les jours de la semaine et le samedi, au besoin.

L'interview de la population active se divise en deux parties : le dossier du ménage (également appelé dossier de contrôle ou liste du ménage) qui comprend des détails sur les membres du ménage et des données démographiques; et le questionnaire de la population active, rempli pour chaque membre du ménage admissible, et comprenant des détails sur les activités de la population active au cours de la semaine précédente. Bien que le texte des deux méthodes d'interview ait été identique à celui de l'EPA, l'instrument ITAO personnalisait automatiquement le texte des questions d'après les réponses aux questions précédentes. De plus, la méthode ITAO comprenait la ramification automatique, l'investigation obligatoire, la révision ainsi que la planification automatique des appels et la gestion de l'échantillon. Il est à noter que les interviews réalisées selon la méthode ITAO ne comportaient pas de questionnaires sur papier, lesquels auraient pu servir de soutien en cas de défectuosité du système. On avisait plutôt les répondants qu'il y avait un problème technique et qu'on les rappelerait aussitôt que possible.

Les échantillons de la méthode ITAO et de la méthode du questionnaire sur papier ont été traités à l'aide du système existant de l'EPA. Pour les deux méthodes, on a fait appel à la révision et à l'imputation, au codage automatisé des branches d'activité et des professions ainsi qu'aux modules de tabulation.

3.0 TAUX DE RÉPONSES

L'ITAO peut influer sur les taux de réponses en raison de la capacité à contacter les ménages admissibles à l'enquête ou de son impact sur l'interaction enquêteur-répondant. Les taux de réponses ont été calculés en fonction de tous les ménages admissibles identifiés pendant l'échantillonnage de la composition téléphonique aléatoire effectué au cours du mois précédent leur introduction dans l'enquête. Le tableau 1 présente l'ensemble des taux de réponses avec la statistique-t pour les différences entre les deux méthodes.

Tableau 1

Taux de réponses par méthode d'interview

	ITAO	Questionnaire sur papier	Statistique-t
Complet	84.3	86.3	-2.05*
Refus	7.8	8.0	-0.17
Non contacté/Absent	7.6	4.8	5.18*
Autre	0.3	0.9	-2.83*

* indique une différence significative à un niveau de 5%.

Dans l'ensemble, la méthode ITAO a donné un taux de réponse inférieur à celui de la méthode faisant appel aux questionnaires sur papier. La différence était presque entièrement attribuable à la difficulté à contacter les ménages, par opposition aux refus. Pour les deux méthodes, 92.1 pour cent des ménages ont répondu au moins une fois pendant la période de couverture de l'enquête, ce qui révèle que l'ITAO n'a pas produit davantage de non-réponses.

3.1 Refus

La majeure partie de l'information contenue dans les publications ayant trait à l'impact de l'ITAO sur les refus est peu concluante. Il semble que le manque de connaissance et de confiance des enquêteurs relativement au système ITAO ait pu avoir un effet (Nicholls, 1978). Toutefois, très peu de données permettent de savoir si, dans le cas où les enquêteurs sont également compétents dans les méthodes d'interviews conventionnelles et de l'ITAO, l'ITAO augmente ou diminue le taux de refus. Les chiffres du tableau 1 n'indiquent aucune différence importante entre les deux méthodes de collecte de données. Bien que nous nous soyons demandé si l'interview ITAO plus structurée pouvait provoquer davantage de refus au cours des premières minutes critiques de l'interview, cela ne s'est pas produit.

Le fait que les enquêteurs aient indiqué aux répondants qu'ils allaient enregistrer leurs réponses sur ordinateur ou encore le bruit du clavier ne semblent pas avoir influé sur leur consentement à collaborer. Bien que d'un mois à l'autre, on ait enregistré une certaine variation dans les taux de refus, il n'y a pas eu de différence marquée dans les huit mois de collecte de données. Il semble que l'ITAO n'ait eu aucun impact sur le taux de refus de l'enquête.

3.2 Contact avec les ménages

La documentation semble indiquer que la planification des appels dans un système ITAO devrait améliorer la capacité à contacter les ménages, grâce à une logique imposée à la fréquence et à la synchronisation des appels pendant une période d'enquête fixe (Groves et Nicholls, 1986). Selon Weeks (1988), la planification automatisée des appels peut éliminer les erreurs de l'enquêteur et faciliter l'utilisation de règles élaborées sur les protocoles d'appels, règles que l'on ne pourrait appliquer avec un système manuel traditionnel.

L'algorithme de planification des appels pour l'échantillon de l'ITAO a fait usage (1) du meilleur temps pour appeler, évalué d'après les mois précédents, (2) des résultats des appels infructueux, (3) du moment où les appels ont été effectués, et (4) des renseignements provenant de l'interview pour déterminer la priorité des appels pendant une période déterminée. La même information était disponible pour les interviews avec les questionnaires sur papier, mais rien n'a été tenté officiellement en vue d'établir la priorité des cas. Les superviseurs ont plutôt assigné un groupe de cas à chaque enquêteur en les laissant classer ces cas à leur gré.

Dans les non-répondants, les ménages non contactés comprenaient les ménages n'ayant jamais été contactés ainsi que les ménages rejoints mais qui ne pouvaient réaliser l'interview à ce moment-là. Comme l'indique le tableau 2, l'élément non-contact était supérieur pour l'ITAO la plupart des mois, bien que la différence n'ait pas été importante au cours des trois derniers mois de l'enquête. Ce taux supérieur de non-contact semble être attribuable à plusieurs facteurs : durée de l'interview, temps d'arrêt des ordinateurs, et expérience et recrutement des enquêteurs.

Tableau 2

Taux de réponses et taux de non-contacts par mois et par méthode d'interview

Mois	Taux de réponses			Taux de non-contacts		
	ITAO	Questionn. sur papier	Stat.-t	ITAO	Questionn. sur papier	Stat.-t
Février	84.3	84.2	0.04	9.4	5.1	2.30*
Mars (1)	80.2	87.7	-3.99*	12.9	4.3	5.99*
Avril	82.7	83.7	-0.58	9.0	5.6	2.75*
Mai	84.1	87.7	-2.40*	7.9	4.2	3.53*
Juin	84.6	89.5	-3.27*	5.9	2.9	3.37*
Juillet	83.6	84.7	-0.74	7.8	6.7	0.96
Août	84.9	83.8	0.63	6.5	6.5	0.01
Septembre	88.9	87.9	0.68	3.5	3.3	0.20
Ensemble des mois	**84.3**	**86.3**	**-2.05***	**7.6**	**4.8**	**5.18***

(1) La semaine du mois de mars consacrée à l'interview selon la méthode ITAO tombait pendant la semaine de congé des élèves.
* indique une différence significative à un niveau de 5%.

Les interviews de l'ITAO prenaient environ 20 pour cent plus de temps que les interviews faisant appel aux questionnaires sur papier. Par conséquent, au cours des périodes d'interviews les plus fructueuses, les enquêteurs pouvaient réaliser davantage d'interviews avec questionnaires sur papier parce qu'ils pouvaient contacter davantage de ménages. Par exemple, les lundis étaient à peu près également fructueux pour les deux méthodes sur le plan du nombre d'interviews réalisées par rapport au nombre de ménages rejoints et pourtant, 13 pour cent plus d'interviews avec questionnaires sur papier ont été réalisées. Plus tard dans la semaine, tandis que la productivité diminuait pour les deux méthodes, l'échantillon ITAO de ménages à appeler était plus important. Bien qu'avec le temps, la différence dans la durée de l'interview entre les deux méthodes soit demeurée assez constante, la durée moyenne de l'interview a considérablement diminué après les trois premiers mois : elle est passée de 11.90 minutes à 7.95 minutes pour la méthode ITAO et de 9.99 minutes à 6.27 minutes pour la méthode faisant appel aux questionnaires sur papier. Cela était principalement attribuable au fait que les enquêteurs étaient devenus plus familiers avec l'enquête et plus confiants dans leurs rapports avec les répondants. Comme la durée de l'interview ITAO a diminué en-deça des huit minutes, le nombre de ménages pouvant être contactés pendant les périodes plus productives a augmenté.

Les temps d'arrêt des ordinateurs ont aussi contribué à augmenter le taux de non-contacts pour la méthode ITAO. Traditionnellement, les lundis, mardis et mercredis constituaient les journées d'interviews les plus productives et il était difficile de re-contacter les menages manqués pendant cette période. Notre programme ITAO fonctionnait sur mini-ordinateur partagé avec d'autres utilisateurs et il y a eu des temps d'arrêt pour plusieurs mois. Pendant un mois type, il fallait approximativement 300 heures d'interviews pour couvrir l'échantillon. Au cours d'un mois, nous avons perdu quatre heures d'interviews un lundi, lorsqu'il y avait huit enquêteurs en service; cette perte équivaut à environ 10 pour cent du temps total d'interview. Un plus grand nombre d'appels le vendredi ne peut compenser une telle perte le lundi.

La dotation en enquêteurs constitue le dernier facteur ayant contribué à la non-réponse. Bien que l'on ait prévu le même nombre d'enquêteurs par méthode, tout au long du test, il y a eu moins d'heures travaillées avec la méthode ITAO qu'avec la méthode du questionnaire sur papier pendant les quarts productifs. On attribue ce fait aux retards et à la maladie. Comme dans le cas des temps d'arrêt des ordinateurs, cette perte en heures d'interviews pendant les périodes de pointe a eu un effet sur le nombre de ménages pouvant être rejoints.

4.0 QUALITÉ DE L'INTERVIEW

Les deux prochaines sections portent sur deux éléments de la qualité des données produites par la méthode ITAO et par la méthode du questionnaire sur papier. Le premier traite de la non-réponse et des vérifications de l'uniformité en établissant une comparaison avec le taux d'erreurs. Le deuxième compare la capacité de coder les réponses avec une série de questions ouvertes sur la branche d'activité et la profession.

4.1 Non-Réponses et erreurs d'uniformité

On attribue fréquemment à la méthode ITAO l'avantage du contrôle rigide du processus d'interview (Nicholls et Groves, 1986). L'enquêteur doit suivre l'ordre des questions et enregistrer une réponse pour chaque question qui s'applique au répondant. En théorie, cela empêche les enquêteurs de sauter des questions mais cela n'assure pas de réponses significatives. Bien qu'il soit possible d'éliminer les réponses telles que "ne sait pas" et "refus" en définissant les réponses acceptables, dans la pratique, l'ITAO admet

ces réponses dans la même mesure que les questionnaires sur papier. Dans une étude, où l'on utilisait le même questionnaire pour la méthode ITAO et pour celle du questionnaire sur papier, Groves et Mathiowetz (1984) ont découvert que les niveaux de non-réponses (combinaison des "ne sait pas", "refus" et "pas certain") étaient à peu près les mêmes pour une série de questions relatives à la démographie et aux revenus.

L'effet le plus important du contrôle rigide du processus d'interview réside peut-être dans le contrôle automatique des séquences de questions. Ceci serait particulièrement vrai pour les applications supposant des instructions de ramifications complexes où le fardeau de la mise en application de la logique du questionnaire ne repose pas sur l'enquêteur. Dans la même étude, Groves et Mathiowetz ont découvert que, pour une série de 28 questions nécessitant une ramification complexe, 1.8 pour cent des données ITAO entrées étaient inconciliables, tandis que 8.8 pour cent des réponses obtenues par une autre méthode contenaient des erreurs d'uniformité.

Le système de traitement des EPA contrôle la validité de même que l'uniformité et produit un rapport de validation. Le rapport fait état des divergences ou du taux d'échec de la validation, c'est-à-dire le pourcentage de réponses inconciliables ou de blancs (ce qui comprend les " ne sait pas" et les "refus"). Ces taux reflètent le total d'erreurs provenant de toutes les sources pendant la collecte et la saisi des données. Bon nombre de sources d'erreurs sont propres aux deux modes, par exemple, la mauvaise compréhension des questions et des erreurs dans l'enregistrement des réponses. Toutefois, l'ITAO et la méthode du questionnaire sur papier constituent des méthodologies de collecte différentes pouvant entraîner des erreurs différentes. Avec la méthode ITAO, les erreurs peuvent être attribuables au matériel ou à la programmation tandis que la méthode du questionnaire sur papier suppose une période distincte de saisie des données pendant laquelle peuvent se produire les erreurs. Comme nous l'avons mentionné plus tôt, l'utilisation de l'ordinateur dans la méthode ITAO devrait réduire ou éliminer plusieurs types d'erreurs. La comparaison de la somme des erreurs attribuables à la collecte et à l'entrée des données illustre l'impact des différentes méthodologies. Le tableau 3 présente les taux de divergences par mois pour le dossier du ménage et le questionnaire sur la population active.

Tableau 3

Taux de divergences par méthode d'interview

Mois	Dossier du ménage		Questionnaire sur la population active		Taux global	
	ITAO	Questionnaire sur papier	ITAO	Questionnaire sur papier	ITAO	Questionnaire sur papier
Février	0.7	1.0	2.0	4.0	1.5	2.7
Mars	0.5	0.9	2.6	7.7	1.7	4.7
Avril	0.4	0.7	1.7	6.3	1.2	3.8
Mai	0.5	0.6	1.8	3.2	1.2	2.0
Juin	0.8	0.7	1.4	3.6	1.2	2.3
Juillet	0.7	0.6	1.6	3.5	1.2	2.2
Août	0.7	0.5	1.4	3.5	1.1	2.1
Septembre	0.7	0.4	1.8	2.9	1.3	1.8
Ensemble des mois	**0.6**	**0.6**	**1.7**	**4.2**	**1.3**	**2.6**

Dans l'ensemble, les taux de divergences pour le dossier du menage sont identifiques. Au cours des quatre premiers mois de l'étude, les taux d'erreurs de l'ITAO étaient inférieurs à ceux de la méthode du questionnaire sur papier mais on a noté l'inverse pendant la dernière moitié de l'étude. Cela semble dépendre des taux d'erreurs différents pour les interviews réalisées au cours des mois subséquents. Au cours de ces mois, une erreur de programmation dans le système ITAO a entraîné une absence de réponses pour l'une des questions, qui s'est reflétée dans les taux d'erreurs du dossier du ménage des derniers mois. Les taux de divergences du dossier du ménage reflètent également un nombre similaire de réponses "ne sait pas" et "refus" pour la méthode ITAO et pour la méthode du questionnaire sur papier. La structure de ramification du dossier du ménage n'est pas complexe. Avec la méthode du questionnaire sur papier, les enquêteurs recueillent les données relatives aux membres du ménage et à la démographie sous forme de tableau, ce qui leur donne une certaine souplesse au niveau de l'ordre dans lequel les données sont recueillies. Pour l'ITAO, il a fallu structurer cette portion de l'interview, et on a simplement classé les questions par ordre. De même, bien que certains contrôles de l'uniformité aient été intégrés au logiciel ITAO, les contrôles d'uniformité plus complexes (par exemple composition de la famille) étaient exclus.

Par contre, la logique de ramification du questionnaire sur la population active est très complexe. La différence dans les taux de divergences pour ce mode reflète en grande partie l'effet du contrôle automatique des séquences de questions dans l'ITAO. Pour la plupart des mois, le taux d'erreurs de la méthode du questionnaire sur papier est deux fois plus élevé que le taux de l'ITAO et dans certains cas, il est trois fois plus élevé. Pour une série de questions du formulaire sur la population active, 23.7% des réponses enregistrées sur le questionnaire sur papier étaient inconciliables comparativement à 4 pour cent pour l'ITAO. Le taux d'erreurs pour l'ITAO n'est pas attribuable au fait que les enquêteurs ne suivent pas les structures de questions de façon appropriée. Les taux de divergences pour les deux méthodes d'interviews reflètent également l'incapacité d'obtenir des réponses significatives.

4.2 Enregistrement des réponses aux questions ouvertes

Bien que le contrôle exercé par l'ITAO puisse nous inciter à croire qu'il est possible de réduire les erreurs au niveau des réponses, la méthode peut cependant augmenter d'autres types d'erreurs. L'augmentation d'erreurs dans les réponses est surtout attribuée aux questions ouvertes où les enquêteurs doivent taper la réponse d'un répondant à une question (Groves et Nicholls, 1986). Si les enquêteurs ne réussissent pas à taper les réponses assez rapidement, il peut se glisser des erreurs typographiques, ou la réponse peut s'avérer incomplète ou inintelligible. Contrairement aux interviews avec questionnaires sur papier où les enquêteurs révisent souvent les réponses aux questions ouvertes une fois l'interview terminée, le logiciel ITAO nécessitait une réponse complète au moment où la question était posée.

Aucune étude antérieure ne compare les réponses textuelles aux questions ouvertes d'une même enquête effectuée avec la méthode ITAO et la méthode du questionnaire sur papier. Toutefois, cette étude offre une mesure permettant d'évaluer l'état complet des réponses à une série de questions ouvertes sur la branche d'activité et la profession. Les réponses à ces questions sont utilisées pour l'affectation des codes relatifs aux branches d'activité et aux professions. Comme première étape du processus, un module de programme tente d'assigner automatiquement les codes appropriés. Le tableau 4 présente les taux de cas automatiquement codés avec la statistique-t pour les différences entre les deux méthodes d'interview.

Tableau 4

Taux de codage par méthode d'interview

Codé automatiquement	ITAO	Questionnaire sur papier	Statistique-t
Février	31.7	29.2	0.88
Mars	33.5	31.6	0.66
Avril	29.7	32.4	-0.63
Mai	34.2	29.9	1.03
Juin	28.1	35.2	-1.75
Juillet	32.2	33.8	0.35
Août	32.9	34.3	-0.32
Septembre	34.0	31.0	0.72
Ensemble des mois	**32.1**	**31.7**	**0.30**

Les taux ne révèlent aucune tendance particulière. Sur le plan statistique, aucune différence significative à un niveau de 5% n'a été enregistrée au cours des huit mois et les taux des deux méthodes étaient à peu près uniformes. On a également étudié le nombre de mots utilisés pour décrire la branche d'activité et la profession. Pendant la durée du test, le nombre moyen de mots utilisés était identique pour la méthode ITAO et pour celle du questionnaire sur papier. Il ne semble pas exister de différence systématique dans la capacité d'enregistrer les réponses aux questions ouvertes.

5.0 ESTIMATIONS D'ENQUÊTES

Les données recueillies selon la méthode ITAO et celle faisant appel aux questionnaires sur papier ont été pondérées afin de permettre la production d'estimations de la population d'après l'enquête. Dans la présente section, on compare ces estimations à celles produites à partir du sondage en cours sur la population active. Toutes les données présentées par l'EPA se rapportent à l'Ontario et excluent les ménages sans téléphone. Cette étude a été conçue dans le but d'établir une comparaison contrôlée des interviews selon la méthode ITAO et celle faisant appel aux questionnaires sur papier et les différences entre les estimations produites par l'EPA et l'une ou l'autre des méthodes s'expliquent sans doute davantage par le contexte du test que par l'impact des méthodologies. Plusieurs différences notables entre les deux enquêtes et l'enquête en cours sur la population active méritent d'être notées : la taille de l'échantillon des méthodes ITAO et du questionnaire sur papier correspond à peu près à 10 pour cent de la taille de l'échantillon de l'EPA; pendant le premier mois, les interviews étaient réalisées par téléphone; les enquêteurs et les superviseurs étaient inexpérimentés et les enquêteurs ne recevaient pas de rapports mensuels sur les échecs relativement aux validations et sur leurs taux de complètement. Les estimations choisies aux fins de comparaison portent sur la taille du ménage, le taux de chômage, le taux de participation et la taille des groupes en ce qui a trait à la branche d'activité et à la profession.

5.1 Taille du ménage

Le tableau 5 indique la taille moyenne des ménages par mois pour l'ITAO, la méthode faisant appel aux questionnaires sur papier et l'EPA. Les différences mensuelles sont minimes et non-significatives sur le plan statistique; toutefois, elles révèlent une tendance constante au cours des mois. La différence dans la taille moyenne des ménages pour l'ITAO et l'EPA est significative pour tous les mois.

Tableau 5

Taille moyenne des ménages par mois pour les ménages avec téléphone - ITAO, questionnaire sur papier et EPA

Mois	ITAO	Questionnaire sur papier	EPA
Février	2.26	2.17	2.14
Mars	2.27	2.15	2.14
Avril	2.23	2.15	2.13
Mai	2.20	2.12	2.13
Juin	2.20	2.17	2.12
Juillet	2.19	2.16	2.12
Août	2.20	2.16	2.13
Septembre	2.18	2.14	2.12
Ensemble des mois	**2.21**	**2.15**	**2.13**

La taille des ménages régulièrement plus élevée de la méthode ITAO peut être attribuable à l'interrogation plus structurée qu'impose la méthode. Chaque méthode de collecte de données oblige les enquêteurs à pousser l'enquête afin de savoir, après avoir demandé la liste des membres du foyer, si on n'a pas oublié de membres du ménage et s'il y a des personnes temporairement absentes. Ces points semblent souvent redondants pour l'enquêteur. Toutefois, comme la méthode ITAO requiert une réponse précise à ces questions, cela diminue vraisemblablement la tendance de l'enquêteur à omettre ces points.

L'information sur les liens de parenté au sein du ménage semble confirmer cette explication. La méthode ITAO a permis de signaler presque deux fois plus de personnes ayant des liens tels que père ou mère, petit-fils, petite-fille, gendre ou bru que ne l'a permis la méthode faisant appel aux questionnaires sur papier et environ 25 pour cent de plus que dans le cas de l'EPA. Sans investigation plus poussée, ces membres du ménage ont tendance à être omis.

La différence dans la taille moyenne des ménages entre l'ITAO et l'EPA équivaut approximativement à trois pour cent, ce qui représente environ la moitié du taux estimé de sous-couverture ou d'écart pour l'EPA pendant la période de cette évaluation. Cela pourrait indiquer que l'interrogation plus structurée de la méthode ITAO peut réduire une partie des erreurs attribuables à l'omission de membres du ménage. L'analyse des données sur les nouvelles interviews permettrait de faire la lumière sur cette question.

5.2 Estimations d'enquêtes sur la population active

Le taux de chômage et le taux de participation représentent deux des plus importantes statistiques provenant de l'EPA. Les estimations de ces deux taux tirées des enquêtes selon la méthode ITAO et celle du questionnaire sur papier suivaient à peu près la même tendance que celle de l'EPA pour la période du test. Les estimations faites à partir des deux méthodes de test sont beaucoup plus irrégulières d'un mois à l'autre en raison principalement des tailles réduites d'échantillons.

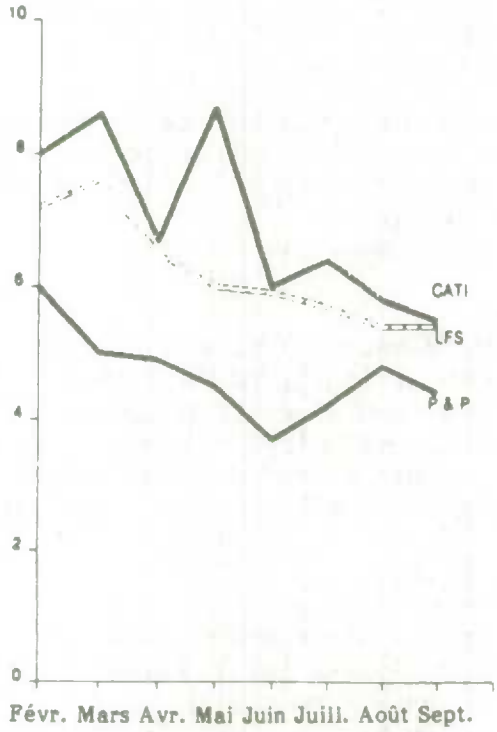
Les estimations du taux de chômage (graphique 1) tirées de la méthode de collecte ITAO étaient toutes supérieures aux estimations provenant de la méthode du questionnaire sur papier. Ces différences étaient considérables en mai, juin et juillet.

Les résultats des deux méthodes de test différaient considérablement de ceux de l'EPA pour un mois : le mois de mai pour l'ITAO, et le mois de juin pour la méthode du questionnaire sur papier.

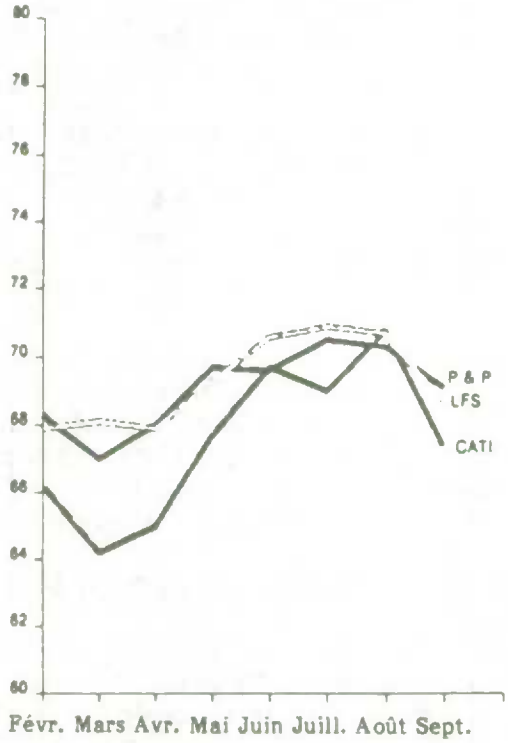
La raison de ces différences n'est pas évidente. L'interrogation plus structurée et le taux d'erreurs plus bas ou le taux plus élevé d'absences de réponses de la méthode ITAO, ou l'impact de la semaine de l'enquête dans ces mois sont des explications possibles. Les différences entre les méthodes ITAO et du questionnaire sur papier ne sont probablement pas attribuables aux enquêteurs étant donné qu'ils ont travaillé avec les deux méthodes, ni à l'échantillon qui provenait de la même source et était assigné au hasard à chacune des méthodes.

Le graphique 2 compare les estimations des taux de participation provenant des trois sources. On n'enregistre aucune différence significative entre ces estimations. On a également comparé les estimations de la taille des regroupements de classification des principales branches d'activité et de professions. Là non plus, il n'y avait pas de différence notable entre les estimations provenant des trois sources.

Bien que les différences entre les trois sources de données ne soient pas significatives pour la plupart des estimations découlant de l'enquête, certaines caractéristiques (taux de chômage et taille des ménages) semblent avoir été touchées par la méthodologie de collecte de données. Les plans en vue d'introduire la collecte assistée par ordinateur pour les enquêtes avec d'importantes séries de données, telle l'EPA, nécessiteront une surveillance étroite et devront être éprouvés.



Graphique 1
Taux de chômage par mois pour les ménages avec téléphone - enquête ITAO, questionnaire sur papier et EPA.



Graphique 2
Taux de participation par mois pour les ménages avec téléphone - enquête ITAO, questionnaire sur papier et EPA.

6.0 COÛT

La méthode ITAO est justifiée en partie par le fait qu'elle combine la saisie des données et les interviews, ce qui devrait permettre de réduire les coûts. Toutefois, le problème s'avère plus complexe que cela.

Certains chercheurs se demandent si l'ITAO augmentera la durée des interviews et par conséquent, leur coût. Harlow, Rosenthal et Ziegler (1985) ont constaté que les entrevues de l'ITAO étaient environ 14 pour cent plus longues. D'autres auteurs ont fait remarquer que les enquêteurs habitués à la méthode du questionnaire sur papier commencent à poser la question suivante pendant qu'ils enregistrent la dernière réponse et qu'ils peuvent inscrire plus rapidement les réponses aux questions ouvertes (Nicholls et Groves, 1985; Groves et Mathiowetz, 1984). Par ailleurs, l'ITAO réduit le volume d'écriture et la supervision (Coulter, 1985) grâce à la planification des appels. Cependant, l'effet net de l'ITAO n'est pas clair.

Dans cette évaluation, les coûts englobés sont ceux qui ont été encourus une fois que l'infrastructure de l'enquête a été mise en place. Ils comprennent les salaires des enquêteurs et des superviseurs, les frais d'interurbains, les coûts d'utilisation de pour la fonction d'interview a fonction d'interview de l'ITAO et la saisie des données pour les interviews faisant appel aux questionnaires sur papier. Les coûts ont été analysés en fonction du nombre de ménages contactés et du nombre d'interviews complétées. Comme on peut le voir dans le tableau 6, dans l'ensemble, les différences sont minimes mais la méthode ITAO semble légèrement moins coûteuse pour chaque ménage contacté et légèrement plus coûteuse pour chaque interview complétée.

Tableau 6

Coût unitaire par méthode d'interview

	ITAO	Questionnaire sur papier	Statistique-t
Enquêteurs	2.87	2.74	+ 5
Supervision	0.63	0.73	-16
Saisie des données/Ordinateur	0.54	0.81	-33
Téléphones*	1.52	1.40	+ 9
Total par ménage	5.55(7,076)	5.68 (6,934)	- 2
Total par interview complétée	6.63(5,963)	6.58 (5,986)	+ 1

* Estimations pour les mois de février, juillet, août et septembre.

Comme le laisse supposer la littérature la supervision est moins coûteuse dans le contexte ITAO. Les superviseurs consacrent moins de temps à la planification des appels ainsi qu'au travail connexe de manipulation et de contrôle des documents. Les témoignages des superviseurs ayant trouvé leur charge de travail moins lourde pendant la semaine consacrée aux interviews selon la méthode ITAO confirment cet état de fait. L'économie pouvant être réalisée en matière de supervision peut être sous-estimée en raison de la faible ampleur de chaque enquête mensuelle.

L'information sur les enquêteurs, donnée au tableau 6, est moins simple. Les interviews complétées de la méthode ITAO duraient en moyenne deux minutes de plus. Les données présentées dans la documentation (Harlow, 1985) le confirment. Le coût

d'interview par ménage contacté était pratiquement égal pour la méthode ITAO et celle faisant appel aux questionnaires sur papier, même si l'interview de l'ITAO était plus longue. Cela s'explique par le fait qu'avec la méthode faisant appel aux questionnaires sur papier, les enquêteurs passent plus de temps entre les appels à réviser le document qu'ils viennent de remplir et à déterminer leur prochain appel.

La méthode ITAO élimine le coût distinct de saisie des données; toutefois, l'utilisation partagée d'un mini-ordinateur par les enquêteurs occasionne des frais dans beaucoup d'applications courantes de la méthode ITAO. Ces coûts, comparés à ceux de l'entrée de données traditionnelle de la méthode du questionnaire sur papier se sont avérés considérablement inférieurs. En dernier lieu venaient les frais de téléphone qui, comme on s'y attendait, étaient plus élevés en raison de la longueur de l'interview. Le coût par interview réalisée indique que la méthode ITAO est légèrement plus coûteuse. Le coût par ménage échantillon semble à peu près équivalent uniquement en raison du fait qu'une plus grande proportion des ménages contactés dans le cadre de la méthode ITAO n'a pas produit d'interviews.

On a laissé entendre que les enquêteurs utilisant la méthode du questionnaire sur papier peuvent inscrire les réponses aux questions ouvertes plus rapidement qu'avec la méthode ITAO et que cela peut influer sur le coût des interviews (Nicholls, 1978). Cela ne semble pas avoir été le cas dans cette étude. Les réponses aux questions ouvertes comprenaient les noms des membres du ménage de même que des descriptions de la branche d'activité et de la profession. Dans le cadre d'une tentative non officielle pour mesurer ce facteur, nous avons contrôlé les interviews et les réponses écrites à la main sur un questionnaire pendant qu'un enquêteur ITAO introduisait les mêmes renseignements par clavier. Il était rarement possible d'égaler la vitesse d'entrée de la méthode ITAO sauf lorsque l'enquêteur demandait des éclaircissements ou faisait épeler un mot. La plupart des enquêteurs engagés pour le projet avaient une certaine expérience au clavier et ce facteur explique certainement la vitesse d'introduction de l'information. D'autres ont été invités à utiliser un programme d'apprentissage de frappe machine.

Nous avons également remarqué que les enquêteurs se sont vite familiarisés avec l'ordre des questions, même lorsqu'ils utilisaient la méthode ITAO, et qu'ils commençaient à poser la question suivante tout en inscrivant la réponse à la question précédente. Ce facteur, que l'on ne retrouve probablement pas dans les enquêtes non répétitives, a certainement été déterminant pour cette application.

7.0 ENQUÊTEURS

Les enquêteurs engagés pour le projet ITAO n'avaient pas d'expérience. En mai et juin 1987, le personnel comptait onze enquêteurs et deux superviseurs, mais en raison du roulement et de besoins moindres, en septembre, le personnel a été réduit à huit enquêteurs et à un superviseur. Tous les enquêteurs étaient bilingues.

La responsabilité principale de l'enquêteur au téléphone consiste à établir un rapport professionnel, amical et de collaboration avec le répondant de façon à obtenir les renseignements requis. Il en est de même avec la méthode ITAO; toutefois, l'imposition de la technologie signifie que l'enquêteur doit avoir d'autres compétences. La majeure partie de l'information contenue dans cette section a été obtenue en écoutant, en surveillant et en observant les enquêteurs pendant les huit mois de l'enquête.

La plupart des enquêteurs savaient comment effectuer la saisie de données ou taper à la machine avant d'être engagés pour ce projet. Ceux qui ne savaient pas taper étaient

beaucoup plus lents pendant la période de formation et ont dû recevoir une formation supplémentaire pour améliorer leur doigté avant de commencer les interviews. Après une formation minimale, ils étaient en mesure de fournir un rendement acceptable.

Les enquêteurs doivent être à l'aise avec l'outil qui leur est fourni pour l'interview. Être assis devant un terminal constituait une nouvelle expérience pour bon nombre d'entre eux. Il a fallu quelques mois avant que les enquêteurs puissent se concentrer sur l'interview sans être distraits par ce qui allait se produire lorsqu'ils appuyaient sur une touche ou par ce qu'il fallait faire en cas d'erreur.

Après les premier, troisième et dernier mois d'interviews, on a recueilli les témoignages des enquêteurs. Le changement d'attitude des enquêteurs est probablement l'élément le plus intéressant qui soit ressorti de ces témoignages. Après le premier mois d'interviews, 80 pour cent des enquêteurs préféraient la version de l'enquête faisant appel aux questionnaires sur papier. Au troisième mois, c'était le contraire et 80 pour cent préféraient la méthode ITAO. Bien que nous ayons escompté que les enquêteurs demeureraient fidèles à la méthode ITAO une fois qu'ils en connaissaient les avantages, ceux-ci étaient légèrement plus portés vers la méthode du questionnaire sur papier à la fin de la période d'interviews. Tout au long des tests, les superviseurs se sont montrés très enthousiastes au sujet de la méthode ITAO. En effet, cette méthode réduisait considérablement leur travail et leur permettait de consacrer plus de temps à la surveillance et à d'autres tâches.

Les enquêteurs appréciaient certains des avantages de l'ITAO tels que l'absence de formulaires de papier, la planification automatique des interviews et le contrôle de la ramification du questionnaire. Cependant, cela ne les a pas empêchés de noter certains inconvénients, à savoir la difficulté d'apporter des corrections, la perte de temps résultant des appels non fructueux et le manque de contrôle sur l'interview. Ce sentiment d'impuissance a été le problème le plus fréquemment soulevé. Les enquêteurs jetaient le blâme sur "l'ordinateur" pour la plupart des problèmes, même ceux qu'ils créaient eux-mêmes.

Les enquêteurs trouvaient donc la semaine consacrée à la méthode ITAO plus stressante. Les rapports des heures travaillées indiquent qu'ils travaillaient moins d'heures pendant les interviews ITAO. D'autres discussions avec certains enquêteurs ont révélé qu'il était plus probable qu'ils ne se présentent pas lors d'un quart d'interview pendant une semaine consacrée à l'ITAO que pendant la semaine de la méthode du questionnaire sur papier. Le niveau élevé de stress avec la méthode ITAO pourrait finalement entraîner un plus gros roulement du personnel jusqu'à ce que celui-ci soit plus à l'aise avec les techniques. Il serait peut-être nécessaire de réduire les quarts ou d'accorder des pauses plus fréquentes.

Au cours du test, les enquêteurs ont pris connaissance du système et du questionnaire ITAO et en sont devenus très critiques. Le temps de réponse - le temps requis pour l'affichage après l'introduction d'une réponse -constituait une source d'irritation pour les enquêteurs. Plus ils se familiarisaient avec la méthode et plus ils prenaient de la vitesse, plus ils voulaient que le temps de réponse soit instantané. Bien que le temps de réponse ait été rapide, une seconde ou moins, les enquêteurs trouvaient que ce court délai interrompait le déroulement de l'interview. Notre temps de réponse était limité non pas par le logiciel, mais par le taux de transmission du terminal. Le matériel empêchait tout accroissement du taux pendant la production bien que lors d'essais, nous ayons augmenté la vitesse de transmission de 4,800 à 9,600 bauds, soit une amélioration considérable. Idéalement, le temps de réponse devrait satisfaire les exigences de l'enquêteur le plus rapide. Les enquêteurs ont également critiqué les écrans

qui ne faisaient pas progresser l'interview, n'affichant que des instructions ou des introductions. Bon nombre de ces écrans ont été raccourcis ou simplement éliminés pendant la durée du test de façon à améliorer le déroulement du questionnaire.

8. CONCLUSION

Nos données ont en majeure partie confirmé les résultats d'autres investigations sur la méthode ITAO ainsi que l'optimisme qui a entraîné la prolifération des systèmes ITAO dans le domaine des enquêtes statistiques.

L'impact majeur de l'ITAO s'exprime en termes de qualité des données. Pour les besoins de cette évaluation, on a transformé un questionnaire sur papier en logiciel ITAO en y apportant très peu de modifications. Même dans cette situation, l'ITAO a réduit le taux d'erreurs par article de 50 pour cent pour l'ensemble des questionnaires et de deux tiers pour le plus complexe des deux questionnaires de l'enquête. Dans l'EPA, chaque mois est traité de façon indépendante et l'on ne tient pas compte des données du mois précédent sauf pour aider les enquêteurs. Les enquêtes qui utilisent des données antérieures ou font des vérifications d'après l'information historique présentent beaucoup plus d'avantages.

Dans notre évaluation, le taux de réponses avec la méthode ITAO était inférieur. Il n'y avait pas de différences en ce qui a trait au taux de refus mais le taux de non-contacts était plus élevé jusqu'aux trois derniers mois de la période d'interview. Cependant, cela semble s'être répercuté davantage sur le coût que sur la qualité. Les interviews prenaient plus de temps avec la méthode ITAO, mais à mesure que la durée des interviews diminuait et que les enquêteurs devenaient plus expérimentés, la différence dans le taux de non-contacts n'était plus évidente. L'interview étant légèrement plus longue, on pourrait remédier à ce problème en augmentant le nombre d'enquêteurs.

Notre étude a démontré que la méthodologie ITAO de collecte de données aura probablement certaines répercussions sur les estimations d'enquêtes. Cela peut être attribuable à la réduction des erreurs ou à l'interview plus structurée de la méthode ITAO. L'introduction de la méthode ITAO dans un programme d'enquête continu nécessitera plus de tests et une surveillance plus étroite.

Les données sur les coûts montrent que la phase d'interview est probablement plus coûteuse avec la méthode ITAO, les interviews étant légèrement plus longues. Toutefois, on peut réaliser des économies au niveau de la supervision et en combinant la collecte et la saisie des données. La qualité accrue pourrait entraîner des économies lors du traitement postérieur à l'enquête. On devra recueillir davantage d'information sur le coût avant de tirer les conclusions finales. L'impact global sur les enquêtes individuelles peut dépendre de leur importance, de leur durée et du nombre de superviseurs par rapport au personnel d'enquête.

Outre les techniques d'interviews normales, la méthode ITAO exige des enquêteurs certaines compétences au niveau de la frappe et les enquêteurs qui n'en possèdent pas peuvent facilement acquérir les compétences minimales requises. Les enquêteurs n'ont pas eu de mal à percevoir les avantages et les inconvénients de la méthode ITAO : elle élimine un bon nombre de tâches banales, toutefois certaines interventions telles que la modification de réponses s'avèrent plus difficiles. Les superviseurs et le gestionnaire de l'enquête ont trouvé que la méthode ITAO présentait des avantages aux chapitres de la gestion de l'enquête et du contrôle de l'échantillon.

RÉFÉRENCES

CATLIN, GARY, and INGRAM, SUSAN (1988), "The effects of CATI on Costs and Data Quality: A comparison of CATI and Paper Methods in Centralized Interviewing," in Telephone Survey Methodology, Robert G. Groves, et al, (editors), New York: John Wiley and Sons. (Forthcoming)

CATLIN, GARY, INGRAM, SUSAN, and HUNTER, LECILY (1988), "The Effect of CATI on Data Quality: A Comparison of CATI and Paper Methods", Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census. (Forthcoming)

GROVES, ROBERT, and MATHIOWETZ, NANCY (1984), "Computer-Assisted Telephone Interviewing: Effects on Interviewers and Respondents", Public Opinion Quarterly, 48(B), 356-359.

GROVES, ROBERT, and NICHOLLS, WILLIAM II (1986), "The Status of Computer-Assisted Telephone Interviewing Part II - Data Quality Issues," Journal of Official Statistics, 2(2), 117-134.

HARLOW, BERNARD, ROSENTHAL, JEANNE, and ZEIGLER, REGINA (1985), "A Comparison of Computer-Assisted and Hard Copy Telephone Interviewing," American Journal of Epidemiology, 112, 335-340.

HOUSE, CAROL (1984), "Computer-Assisted Telephone Interviewing on Cattle Multiple Farm Survey", Staff Report NO. 82, Statistical Research Division, U.S. Department of Agriculture.

NICHOLLS, WILLIAM II (1978), "Experiences with CATI in a Large-Scale Survey", Proceedings of the American Statistical Association, Section on Survey Research Methods, 9-17.

NICHOLLS, WILLIAM II, and GROVES, ROBERT (1986), "The Status of Computer-Assisted Telephone Interviewing: PartI - Introduction and Impact on Cost and Timeliness of Survey Data", Journal of Official Statistics, 2(2), 93-115.

STATISTICS CANADA (1984), "Telephone Ownership in Canada" Unpublished report. Prepared by Gary Catlin, G. Hussain Ghoudhry, and Hank Hofmann.

WEEKS, MICHAEL (1988), "Call Scheduling with CATI: Current Capabilities and Methods," in Telephone Survey Methodology, Robert G. Groves, et al. (editors), New York: John Wiley and Sons. (Forthcoming)

DEUX NOUVELLES TECHNOLOGIES DE COLLECTE EXPÉRIMENTÉES À L'INSEE

C. BERNARD,

Institut national de la statistique et des études économiques

RESUME

1 - LE VIDEO-QUESTIONNAIRE

En 1986, l'INSEE a réalisé une expérience de collecte de l'enquête mensuelle de conjoncture dans l'industrie à partir de claviers-écrans (MINITEL) installés dans les entreprises.

Cette expérimentation nous a permis de connaître les conditions d'emploi de cette nouvelle technique, appelée Vidéo-questionnaire, et d'envisager sa mise en production sur une enquête possédant les caractéristiques les plus appropriées à cet outil.

2 - LA COLLECTE ASSISTEE PAR MICRO-ORDINATEUR PORTABLE

Les recherches entreprises depuis quelques années par l'INSEE dans ce domaine ont abouti en 1987 à la réalisation d'un test sur l'enquête quadrimestrielle de conjoncture auprès de 250 ménages.

Ce test cherchait à vérifier principalement la bonne acceptation des micros-ordinateurs par les enquêtés et les enquêteurs ainsi que la faisabilité technique.

Les principales conclusions auxquelles nous sommes arrivés sont :

. que le micro-ordinateur a été bien accepté par les enquêtés et les enquêteurs ;

. que le taux de refus, a été sensiblement le même que dans l'enquête classique, et en aucun cas lié à la présence du micro-ordinateur ;

. que les durées d'interview sont comparables à celles du mode classique ;

. que la réalisation technique est possible à des coûts acceptables ;

. qu'une évolution des qualifications est nécessaire.

La prochaine étape sera pour nous, de mettre en place, au début de l'année 1989, un test sur l'enquête emploi en vue de sa rénovation en 1990. Ce test effectué auprès de 2000 ménages devrait permettre de mieux cerner l'impact de l'outil sur la qualité des données.

---

(1) Exposé préparé par Madame BERNARD.
Les points de vue exposés dans ce papier n'engagent que son auteur et ne reflètent pas nécessairement la position de l'INSEE.

IMPACT DES NOUVELLES TECHNOLOGIES SUR LES ENQUETES

Concernant les nouvelles technologies de collecte d'enquête, l'INSEE travaille principalement dans deux directions : l'utilisation du vidéo-questionnaire et celle du micro-ordinateur portable.

Avant d'aborder le deuxième point qui est l'objet principal de l'exposé, je donnerai rapidement quelques informations concernant nos expérimentations du vidéo-questionnaire.

* *
*

A terme, cette technique devrait nous apporter des améliorations dans les délais et la qualité des données collectées. Nos essais ont constitué, jusqu'à présent, à inviter 158 entreprises à répondre à notre enquête mensuelle de conjoncture dans l'industrie à partir d'un clavier-écran installé chez elles et ce, pendant un an. La collecte de cette enquête s'effectue normalement par voie postale. Le clavier-écran disponible dans l'entreprise était un MINITEL, possédant 24 lignes et 40 colonnes ; il est distribué gratuitement par FRANCE TELECOM (ex : PTT) à tous les abonnés du téléphone. Sa fonction première est de remplacer le bottin téléphonique papier pour la recherche du numéro d'un abonné. Ainsi, munie de cet appareil, l'entreprise pouvait, en composant le numéro d'accès au serveur, constitué par un ordinateur de l'INSEE, voir apparaître le menu d'interrogation et répondre à notre enquête. La première phase de ces essais est aujourd'hui terminée : elle nous permet d'affirmer que cette procédure est techniquement possible, mais que sa rentabilité reste encore à démontrer. En particulier, les diminutions de coûts attendus sur les frais postaux ne se sont pas produites. En effet, la fidélisation mensuelle ne s'est pas réellement opérée et une relance systématique par courrier a dû être mise en place. Par ailleurs, un suivi, une assistance et une légère formation des enquêtés paraissent encore nécessaires même si le besoin tend à disparaître au fur et à mesure de l'intégration de cet outil à la vie courante de la population. La mise au point de telles applications est délicate et une mauvaise ergonomie des grilles provoque des abandons et probablement des biais dans les résultats.

Cette technique s'adapte d'autant mieux à une enquête :

- que celle-ci est répétitive ;

- qu'une seule personne peut répondre à toutes les questions ;

- *que le questionnaire est assez simple pour que l'on puisse y répondre en moins de 10 minutes ;*

- *qu'il y a peu de grilles à remplir ;*

- *que les grilles sont de structure homogène ;*

- *que les questions sont qualitatives car une question quantitative peut entraîner des recherches qui risquent de ne pas être faites avec un outil qui privilégie la réponse instantanée ;*

- *que les motivations des enquêtés à répondre à l'enquête sont réelles ;*

- *que la conception du questionnaire prend en compte les caractéristiques de l'outil qui privilégie une forme de dialogue direct, bref et simple.*

C'est pourquoi notre deuxième étape consistera à choisir une enquête présentant les critères décrits ci-dessus et dont l'implantation en vraie grandeur sera très progressive, ceci afin de permettre une observation et une consolidation de ce nouveau mode de collecte au cours du temps.

* *
*

Parallèlement à ces études de Vidéo-questionnaire, l'INSEE mène depuis 1982 des investigations dans le domaine de la collecte assistée par micro-ordinateur portable. Jusqu'en 1986 la plupart de nos essais se sont heurtés à des difficultés techniques de réalisation, dues principalement aux limites des matériels et logiciels.

L'amélioration ces dernières années de la puissance et de la fiabilité des matériels ainsi que les expériences menées dans ce domaine par certains pays, notamment aux PAYS-BAS, nous ont permis de progresser. C'est ainsi qu'en décembre 1987, un test a été organisé sur l'enquête quadrimestrielle de conjoncture auprès de 250 ménages.

OBJECTIF VISE -

Le but de ce test était de mesurer la faisabilité technique et humaine d'une telle opération ainsi que de noter les réactions des enquêteurs et des enquêtés.

Afin de pouvoir comparer cette nouvelle procédure à la collecte classique sur papier, il a été décidé de reproduire à l'identique le questionnaire et les contrôles manuels jusqu'à la phase de saisie incluse.

LES CARACTERISTIQUES DE L'ENQUETE -

L'enquête de conjoncture auprès des ménages comporte environ 240 questions organisées en huit sous-questionnaires (composition du ménage, gros équipement, automobile, logement, téléphone, questions générales, vacances, fin du questionnaire). La partie consacrée aux vacances est sans doute la plus complexe : on cherche à connaître pour chaque membre du ménage ses périodes de vacances en contrôlant la cohérence temporelle et géographique de l'ensemble.

L'échantillon retenu pour le test a été fixé à 250 logements situés dans une seule région française. Les logements tirés étaient voisins, c'est-à-dire géographiquement proches de ceux enquêtés selon la procédure classique. Ainsi les mêmes conditions de collecte (coûts, déplacements, habitat) étaient-elles reproduites. Toutes les strates géographiques étaient représentées avec des proportions équivalentes à leurs poids dans la région. La répartition de l'échantillon au sens des critères catégorie sociale, taille du ménage et habitat a fait apparaître des "quota" très proches de ceux enregistrés dans l'échantillon de référence régional.

L'ORGANISATION DU TEST -

Dix enquêteurs volontaires (huit femmes, deux hommes) représentatifs du réseau d'enquêteurs de la région ont participé aux tests. Ils ont reçu une formation de deux jours destinée à l'apprentissage des manipulations du matériel et à la mise au point définitive du programme informatique d'interview.

Un certain nombre de mesures concernant le déroulement de l'enquête classique avait été effectuées avant le début des tests. Nous voulions en particulier préciser les circonstances habituelles des refus de réponse, la durée de l'enquête, les conditions matérielles de l'entretien (dans l'escalier pour moins d'un % des cas, à l'intérieur du logement, etc...).

Au cours du test, les enquêteurs ont eu à remplir un questionnaire, à la suite des deux jours de formation, puis après chaque interview, et en fin de test. De plus à la fin de chaque entretien les réactions et impressions des ménages sur ce nouveau mode de collecte étaient recueillies. L'exploitation de cet ensemble d'informations nous a permis de tirer le bilan présenté ici.

Pour neuf enquêteurs sur dix le déchargement des questionnaires saisis s'est effectué en milieu et fin d'enquête dans les locaux de l'INSEE. Le matériel du dixième était équipé d'un modem permettant la transmission à distance des questionnaires par ligne téléphonique.

LA REALISATION TECHNIQUE -

Les machines utilisées sur le terrain se composaient de quatre EPSON PX4 dont la capacité mémoire avait été portée à 192 k et de six TOSHIBA 1100+ (MS/DOS) de capacité mémoire 640 k.

La génération du questionnaire a été réalisée en cinq jours sur un micro-ordinateur de bureau MS/DOS compatible IBM/PC supportant le logiciel SIC (Système Intégré de Collecte). Ce logiciel, encore à l'état de prototype, a été développé par l'ORSTOM (Office de la Recherche Scientifique et Technique d'Outre-Mer), organisme chargé d'enquêtes dans les pays en développement. La convivialité de ce produit nous a permis de développer la totalité du questionnaire et les enchaînements requis sans modification majeure. Ainsi, 12 000 lignes de TURBO PASCAL ont été générées qui une fois réduites en langage assimilable par la machine de terrain constituaient un module de 31 k-octets.

B I L A N -

Il est clair que l'objectif visé a été pleinement atteint et nous a permis de mieux cerner les impacts de cette nouvelle technologie.

Ainsi nous avons pu vérifier :

* ***que le micro-ordinateur a été bien accepté par les enquêtés et les enquêteurs.***

En effet, les enquêtés ont peu réagi à la vue du matériel. Toutefois une concentration et une attention plus accrues des enquêteurs sur leur machine peuvent expliquer, dans certains cas, que le dialogue avec l'enquêté ait été moins riche, sans que cela ait nui, selon les enquêteurs, à la qualité de l'enquête. Il est à noter que cette concentration était partagée par les enquêtés, qui selon les enquêteurs, réfléchissaient plus avant de répondre parce que les réponses étaient enregistrées dans la machine.

Les enquêteurs se sont bien adaptés et ont manifesté un grand enthousiasme pour poursuivre dans cette voie. Notons toutefois qu'ils étaient à l'origine volontaires. Ils ont trouvé que les enchaînements automatiques des questions non seulement ne les gênaient pas, mais étaient plus rapides et plus fiables. Ils ont regretté qu'il n'y ait pas eu plus de contrôles, et ont éprouvé des difficultés de lecture sur l'écran entraînant une certaine fatigue visuelle.

Contrairement à ce que nous pouvions penser, six enquêteurs sur dix ne se sont pas plaints du poids du matériel, alors que le TOSHIBA pesait 4,5 kg.

Toutefois, il nous paraît nécessaire que ce poids décroisse, surtout si l'on souhaite étendre l'emploi de ces matériels à des enquêtes telles que les relevés de prix, qui impliquent une grande mobilité et un travail principalement en station verticale.

* ***que le taux de refus, légèrement inférieur à celui de l'enquête classique, n'est en aucun cas lié à la présence du micro-ordinateur.***

Par rapport aux résidences principales il y a eu :

pour le test de décembre 1987	pour l'enquête classique d'octobre 1987
92,6% d'enquêtes acceptées	89,6% d'enquêtes acceptées
3,9% de ménages non contactés	5,1% de ménages non contactés
3,5% de refus	5,3% de refus

Le taux de refus présenté ci-dessus dans le test est global ; la faiblesse de la représentativité des différentes catégories de ménages à l'intérieur de l'échantillon rend peu signifiante la répartition des refus. Cet aspect sera étudié plus finement lors du prochain test sur l'enquête emploi. Notons toutefois que tous les ménages qui ont opposé un refus n'avaient pas encore vu le micro-ordinateur, et que comme dans l'enquête classique, elles l'ont fait pour des considérations d'ordre moral (atteinte à la vie privée) ou politique (situation économique).

* ***que les durées d'interview sont comparables à celles du mode classique.***

La durée d'interview a été sensiblement la même que dans l'enquête habituelle soit 35 minutes en moyenne. Toutefois les enquêteurs ont fait observer qu'ils réalisaient moins d'enquêtes par jour en raison d'une plus grande fatigue et ont relevé qu'il y avait là un impact sur leur rémunération.

Compte-tenu du verrouillage automatique des grilles de réponses à l'extinction de la machine, une économie de dix minutes en moyenne a été réalisée sur chaque questionnaire. En effet, dans la procédure classique ce temps est consacré à domicile à la mise au propre, à la vérification et à la complétude par téléphone des réponses partielles. Six enquêteurs sur dix ont dit avoir été gênés par l'impossibilité de cette reprise sans que l'on puisse toutefois constater de différence significative sur le nombre de réponses partielles.

* ***que la réalisation technique est possible à des coûts d'analyse et de programmation acceptables.***

  - *l'utilisation d'un générateur de questionnaires s'est avérée être une condition préalable pour atteindre des coûts de développement acceptables. Les possibilités du prototype SIC nous ont permis de respecter la complexité du questionnaire et d'autoriser les contradictions des enquêtes lorsqu'il fallait le faire.*

Toutefois, il faut noter que l'option prise de reproduire le plus fidèlement possible la procédure papier n'est pas la meilleure solution. En particulier la présentation des tableaux de séjours et des biens d'équipement aurait été plus simple si les possibilités du matériel avaient été prises en compte.

- *il ressort des comparaisons des deux matériels utilisés que :*

  - *les capacités de l'EPSON PX4 ont été trop limitées en taille d'écran et de mémoire pour notre application et que son système d'exploitation CP/M a été une gêne dans notre organisation. En particulier quelques fonctions de retour et modifications dans certains sous-questionnaires n'ont pu être installées sur cette machine.*

  - *le matériel TOSHIBA 1100+, bien que très lourd, a été mieux accepté, et a mieux répondu à notre attente ; toutefois son écran est difficile à lire.*

  - *L'expérience nous a montré qu'une autonomie d'au moins huit heures doit être requise sur ces machines.*

- *Les communications entre le matériel des enquêteurs et la machine de l'Institut (un micro-ordinateur compatible IBM.PC) ont été fiables et faciles à mettre en oeuvre.*

* ***Qu'une évolution des qualifications est nécessaire.***

- *Les remarques émises jusqu'à présent tendent à montrer que des études sur l'évolution des qualifications requises et la formation des enquêteurs doivent être entreprises.*

- *Les enquêteurs souhaitent acquérir une certaine dextérité de frappe. Dans ce nouveau mode de collecte, ils s'interrogent davantage sur les processus aval de production et proposent spontanément des contrôles supplémentaires.*

- *L'expérience a mis en lumière la nécessité de redéfinir le partage des tâches entre l'enquêteur et l'atelier manuel notamment dans la réception, la gestion, la saisie et le contrôle des questionnaires. Ce poste en atelier appelé "serveur de collecte" requiert, d'un point de vue technique, la mise à disposition d'un logiciel de messagerie couplé à un transfert de fichier.*

Il est à noter qu'en raison de la taille trop faible de l'échantillon la comparaison des données collectées avec celles de l'enquête classique ne peut permettre de conclure significativement à l'absence de biais ou à une différence de qualité.

Le succès de cette opération nous conduit à aller plus loin en mettant en place, dans le cadre de la préparation de la rénovation de l'enquête emploi pour 1990, une expérimentation de collecte assistée par micro-ordinateur qui devrait avoir lieu au début de l'année 1989 auprès de 2 000 ménages (soit 50 aires).

L'objectif de ce nouvel essai sera de mieux cerner l'impact de l'outil sur la qualité des données, et de réaliser l'ensemble des traitements d'apurement et d'exploitation en mettant en place la fonction de "serveur" de collecte.

* *
*

C O N C L U S I O N S -

Les résultats obtenus jusqu'à présent montrent que la technique permet de recueillir directement sur micro-ordinateur, sans risque de rejet, des données d'enquête auprès des ménages. Cette technique modifie les organisations actuelles et requiert en particulier un nouveau partage des tâches entre les enquêteurs et les ateliers manuels. Ce dernier aspect est peut-être un peu différent de ce que j'ai pu lire sur les expérimentations présentées par les Pays-Bas, ceci étant, sans doute, lié à notre organisation.

Cette technique s'adapte d'autant mieux aux enquêtes répétitives présentant des contrôles "historiques", de nombreuses questions filtres et un nombre limité de questions ouvertes.

En fait, pour bénéficier pleinement des avantages d'interactivité et de délai qu'elle offre, il faut tenir compte des contraintes inhérentes à l'outil, en particulier dans la façon de poser les questions de présenter les modalités des réponses, et de les contrôler. Ceci peut paraître contraire à son emploi dans des enquêtes répétitives ou longitudinales déja existantes ; c'est pourquoi, compte tenu du poids du passé, nous nous imposons une étape supplémentaire de contrôle qui devrait permettre de cerner les éventuelles ruptures de série.

* *
*

BIBLOGRAPHIE

- *Compte rendu de mission : 4ème réunion du groupe conjoint sur les systèmes généraux de traitement d'enquêtes - Programme de calcul statistique VOORBURG 8-12 décembre 1986 - INSEE division logiciels 23 décembre 1986 - note n°1263/640 JPG/MaP.*

- *Compte rendu de la réunion du groupe de travail conjoint sur les sytèmes généraux de traitement d'enquêtes - Programme de calcul statistique - Londres 6-10 avril 1987 - INSEE - division logiciels note n°091/640 JPG/MaP.*

- *Réunion sur la méthodologie statistique (17-20 février 1986) point 5 - utilisation de micro-ordinateur portatif au bureau central de statistique des Pays-Bas (14 novembre 1985 CES/AC.48/63) par Micke E.J. BEMELMANS-SPORK, Wouter J. KELLER et Dirk SIKKEL.*

- *Working Party on Electronic Data Processing (16-20 février 1987). Micros and local computers at Statistics Sweden par L. NORDBACK - 24 novembre 1986 (CES/WP.9/229).*

- *35ème session plénière - 15-19 juin 1985. The impact of the developpement of EDP on statistical methodology and survey techniques - report by Statistics Sweden par Lars LYBERG et Bo SUNDGREN CES/587 - 1er avril 1987.*

- *Computer assisted data collection - (Netherlands Central Bureau of Statistics - Department of statistical methods) 5 décembre 1986, BPA n°15035-86-M1.*

- *A test of the Netherlands continous labour force survey with hand-held AE computers : Interview Behaviour and data quality by VAN BASTEL, Frans KERSSEMAKERS, Dirk SIKKEL, BPA n°947-86-M1 - 26 janvier 1987.*

- *The status of computer - Assisted telephone interviewing : Part I - Introduction and impact on cost and timeliness of survey data - Part II - data quality issues : William L. NICHOLS II et Robert M. GROVES - Journal of Official Statistics vol 2 n°2 - 1986, pp 93 - 134.*

- *The first lap-top computer fieldwork project in SSD/OPCS (March 1987).*

- *Observations of prices with hand-held computers Micke BEMELMANS-SPORK, Dirk SIKKEL (septembre 1984) - BPA n° : 13229-84-M1.*

- *BLAISE SYSTEM (Netherlands Central Bureau of Statistics) JG. BETHLEHEM, D. DENTENEER, A.J. HUNDEPOOL, M.H. SCHUERHOFF, W.J. KELLER.*
  - *The BLAISE system for computer - asssisted survey processing - BPA n°14896-86-M5 january 1987 ;*
  - *A new approch to computer assisted survey processing - BPA n°13436-87-M3 - 6 Novembre 1987.*
  - *Automating the data editing process with the BLAISE system - BPA n°14263-87-M3 (24 novembre 1987).*

- *Automation in survey processing - CBS select 4.*

- *Description du logiciel SIC (Société AGERIS) avril 1988.*

A N N E X E S

- ooOoo -

1 - Structure de l'enquête quadrimestrielle de conjoncture.

2 - Caractéristiques des matériels loués pour le test.

3 - Tableaux

4 - Questionnaire de l'enquête de conjoncture auprès des ménages.

STRUCTURE DU QUESTIONNAIRE DE L'ENQUÊTE DE CONJUNCTURE PAR LE LOGICIEL SIC

```
IG   Informations Générales  S
                    CM   Composition du Ménage  S
                         MM   Membres du Ménage  M
                    GE   Gros Équipement  S
                         RE   Refrigerateur  S
                         LL   Machine à Laver le Linge  S
                         TV   Télévision  S
                         LV   Lave Vaiselle  S
                         CS   Congelateur Simple  S
                         RC   Combine R.C.S.  S
                         IA   Intentions d'Achat  S
                    AU   Automobile S
                         V1   Voiture  M
                    LO   Logement Principal  S
                    TE   Téléphone  S
                         US   Usage de Téléphone  S
                    GS   Questions Générales  S
                    VA   Vacances  S
                         IN   Individus du Ménage  M
                              SE   Séjours  M
```

CARACTÉRISTIQUE DES MATÉRIELS LOUÉS POUR LE TEST

Deux types de matériel ont été utilisés au cours de l'enquête:

```
1 -  TOSHIBA 1100+

     Capacité mémoire: 640 K
     Taille écran: 24 lignes x 80 caractères
     Autonomie: 4 H
     Poids: 4,5 kg
     Système d'exploitation: MS DOS
```

```
2 -  EPSON PX4

     Capacité Mémoire: 64 K + extension 128 K
     Taille écran: 8 lignes x 80 caractères
     Autonomie: 8 H
     Poids: 2 kg
     système d'exploitation: CPM version 2.2
```

Quatre enquêteurs ont travaillé avec l'EPSON, les six autres évoluaient avec le TOSHIBA 1100 +

On a enregistré cinq pannes, surtout des problèmes de batterie.

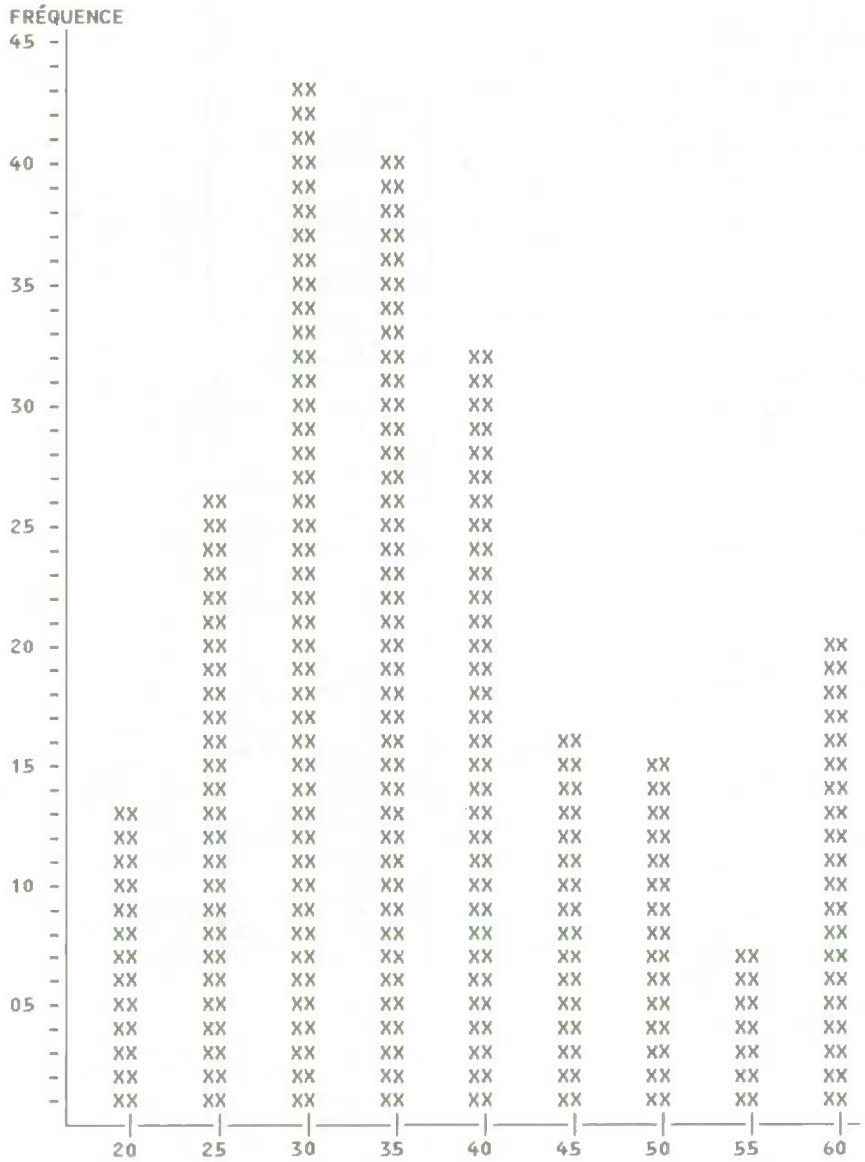
JUDGEMENT COMPARATIF

Judgement comparatif (par rapport au papier) des enquêteurs sur la qualité des données recueillies

Table de Qualité par Matériel

Qualité	Équipment		
Fréquence	EPSON	TOSHIBA	Total
Superieure	37	24	61
Équivalente	40	97	137
Inférieure	5	2	7
Non-Declaré	4	3	7
Total	86	126	212

RÉPARITION DES INTÉRVIEWS SELON LA DURÉE



DURÉE MOYENNE DES INTERVIEWS : 38 Minutes

LES BANQUES INFORMATISÉES DE QUESTIONS: POURQUOI ET COMMENT?

R. P. BAKER,

National Opinion Research Center

INTRODUCTION

Le but de toute enquête est de récolter de l'information sur un sujet d'intérêt. On poursuit ce but en concevant des instruments de collecte de données qui chercheront une grande variété de données très précises, définies et reliées entre elles de façon naturelle. En réunissant toutes ces informations en un même endroit et en étudiant les relations qui les unissent, on espère mieux comprendre les attitudes et les comportements sociaux. Autrement dit, le but d'une enquête est de mettre au point une base de données sur un sujet important.

Le questionnaire d'enquête joue un rôle crucial dans ce processus de mise au point de l'information. D'abord, il définit les questions qu'il faudra poser. Mais ce n'est pas tout: il complète le plan d'échantillonnage en déterminant quelles personnes seront interviewées, et il influence à un certain point les résultats de la recherche en exprimant nos idées préconçues sur la façon dont certains faits et brins d'information entrent en corrélation pour former la connaissance.

Bien sûr, chacun des protagonistes du traitement d'enquête voit le questionnaire d'un oeil différent. Pour le concepteur, le questionnaire définit les données que l'enquête doit récolter. Le concepteur spécifie non seulement les questions qu'il faudra poser en termes précis mais aussi l'éventail des réponses possibles. Il traduit l'intérêt général d'un analyste sur l'état matrimonial d'une population, par exemple, en questions spécifiques sur le nombre de personnes mariées, séparées, divorcées, veuves, n'ayant jamais été mariées, et ainsi de suite.

Pour le concepteur de systèmes d'enquêtes, le questionnaire spécifie les métadonnées, c'est-à-dire, l'information sur les données que l'enquête récoltera. Ces métadonnées nous indiquent à quel type de données s'attendre, l'éventail des réponses possibles, et les règles à suivre pour le traitement des données. Pour le concepteur de systèmes, une enquête est un processus de mise au point d'une base de données et le questionnaire d'enquête est le dictionnaire pour cette base de données.

Enfin, l'analyste des données d'enquête utilise le questionnaire pour planifier l'analyse de la base de données. Le questionnaire décrit pour l'analyste l'information spécifique contenue dans les variables analysées, les liens qui existent entre ces variables et les différentes sous-populations auxquelles elles pourraient s'appliquer.

Le présent exposé met l'accent sur les questionnaires d'enquête, leur conception et leur gestion par ordinateur. Il part du principe que le questionnaire d'enquête est le document clé pour la définition de toute enquête. En utilisant la technologie informatique de façon appropriée, les questionnaires peuvent être conçus beaucoup plus efficacement, et la cohérence, la fiabilité et le degré d'utilité des données d'enquête peuvent être accrus. Ce document a comme point de départ et point d'arrivée les banques de questions, car celles-ci nous mènent vers de nouvelles façons d'utiliser l'ordinateur dans la réalisation des enquêtes et dépasseront leurs limitations actuelles grâce à l'évolution de la technologie informatique et de la recherche sur les techniques d'enquête.

Ce document est divisé en quatre parties. Dans la première, on fait une description générale des banques de questions et on présente l'emploi qui en est fait pour les enquêtes. On regarde ensuite comment cet emploi définit les conditions requises pour obtenir des banques de questions informatisées complètes. Dans la troisième partie, on regarde comment les tendances principales de la technologie informatique et de la recherche en matière de techniques d'enquête continuent de nous mener vers la constitution de banques de questions complexes aux fonctions multiples. Finalement, dans la conclusion, on résume les points essentiels de la discussion.

1. LES BANQUES DE QUESTIONS ET LES ENQUÊTES

L'expression banque de questions évoque des réalités différentes selon les personnes. Sous sa forme la plus simple, la banque de questions n'est qu'un catalogue de questions d'enquête, normalement classées par sujet. Elle peut être sous forme lisible par machine et consultable par ordinateur, ou elle peut être un simple document sur support papier comportant des questions classées par sujet et peut être complété par des renvois sur les différents sujets. Sous sa forme la plus avancée, qui n'a pas encore été atteinte, la banque de questions pourrait se trouver au coeur des fusions les plus complexes et les plus puissantes de la technologie informatique et de la recherche sur les techniques d'enquête.

1.1 Les banques de questions et la conception des questionnaires

Les banques de questions semblent avantageuses aux concepteurs de questionnaires pour deux raisons. D'abord, on suppose que les questions qu'elles contiennent fonctionnent, c'est-à-dire, qu'elles ont été écrites par des experts et sérieusement testées pour démontrer qu'elles obtiennent des informations utiles au cours des enquêtes. Ensuite, elles assurent la standardisation, grâce à laquelle il est plus probable que les nouvelles enquêtes produiront des données comparables à celles récoltées sur un même sujet lors d'enquêtes antérieures.

Un bon exemple de cette simple forme de banques de questions fut récemment publié: The Question Finder (Owen et Mauro, 1988). Cet ouvrage de référence sur support papier comporte au-delà de 5,000 questions d'enquête tirées de divers sondages et enquêtes menés auprès des consommateurs aux États-Unis au cours des 25 dernières années. Les questions sont classées par sujet en 91 catégories, permettant ainsi aux concepteurs de questionnaires de trouver rapidement les questions portant sur les domaines qui les intéressent.

Le concept fondamental de banques de questions est souvent élargi par l'inclusion d'information sur les enquêtes au cours desquelles une question a été utilisée, ainsi que la fréquence avec laquelle les interviewés ont choisi une réponse donnée. La Public Opinion Location Library (POLL) du Roper Center est un exemple; il s'agit d'un système SPIRES[1] d'extraction des données comportant plus de 85,000 questions posées lors d'une grande variété de sondages tels que ceux menés par Roper, Gallup, Harris, NORC, NBC, ABC-Washington Post, et CBS-New York Times (Roper Center, 1987). Chaque entrée (question) présente le texte intégral de la question et de la réponse, la répartition en pourcentage des réponses obtenues, et la description de l'enquête de laquelle la distribution a été tirée, comprenant la population cible, le nombre d'interviews, l'organisme d'enquête, les dates du début et de la fin de l'enquête, et une référence au questionnaire spécifique duquel la question a été tirée. Les utilisateurs de la POLL peuvent consulter sa base de données via un index par sujet, la formulation de la question, l'organisme d'enquête, ou les dates de l'enquête.

La POLL pourrait éventuellement servir à des fins qui dépassent celles du concepteur de questionnaires à la recherche de bonnes questions. La POLL est à la fois une banque de questions classiques et une précieuse source d'information pour toute personne sérieusement intéressée à l'opinion publique aux États-Unis.

---

[1] SPIRES (Standard Public Information Retrieval System) est un système de base de données permettant la mise en mémoire de textes à structure non imposée et l'extraction d'information au moyen d'un index par mot-clé.

1.2 Autres emplois de la banque de questions dans les sondages

Un des principaux sujets traités dans ce document est l'application de la technologie informatique à la mise au point et à l'utilisation des banques de questions: on tente de savoir jusqu'à quel point l'informatique peut mener les banques de questions au-delà de l'emploi qu'en fait actuellement le concepteur de questionnaires, c'est-à-dire de s'en servir comme menu de questions. Des projets de mise au point de logiciels en cours dans des organismes d'enquête partout sur le globe visent à élargir le concept de banque de questions pour atteindre un vaste ensemble d'utilisations potentielles. Cinq de ces nouveaux emplois sont déjà très évidents: (1) la production automatisée de questionnaires, (2) l'évaluation et le contrôle de l'efficacité des questions, (3) la mise en place automatisée de systèmes de saisie des données (ITAO et IPAO), (4) la production de documentation d'enquête, et (5) la mise au point de dictionnaires de données pouvant commander les logiciels d'extraction.

Les questionnaires produits par ordinateur

Imaginons un système automatisé qui permet à un concepteur de questionnaires de lire un catalogue de questions, de choisir les questions qu'il ou elle désire utiliser, et de demander au système de produire un questionnaire sur support papier complètement formaté. Un tel système est certainement possible, surtout pour des questionnaires relativement simples et courts. Mais comme les questionnaires deviennent de plus en plus complexes, la constitution d'un tel système devient plus difficile.

Une principale source de complications est la nécessité de mettre au point un système pour maintenir disponibles les instructions "passez à" d'un questionnaire. Beaucoup de questions qu'on retrouve dans une banque n'auront été posées qu'à un sous-groupe des sujets interrogés lors d'une entrevue ou seulement dans le contexte d'un ensemble de questions reliées entre elles. Ce serait dangereux de les utiliser de façon isolée ou même de juger si elles sont appropriées sans que le contexte des enquêtes dans lesquelles elles ont servi antérieurement n'ait été clairement défini.

Il faut donc commencer à concevoir des banques de questions dans lesquelles on peut mettre en mémoire de l'information contextuelle en plus du texte des questions, des réponses et de la répartition des réponses. Une façon d'atteindre cet objectif est d'introduire dans la base de données la possibilité de reconstituer les instructions "passez à" du questionnaire original dans lequel les questions ont été utilisées. Imaginons qu'un utilisateur passe en revue un ensemble de questions sur un sujet précis. Il choisit une question qui lui semble intéressante et peut obtenir de l'information supplémentaire lui indiquant où elle est apparue au cours de diverses enquêtes, chaque fois, présentant une répartition différente des réponses et précédée de questions différentes. À ce stade, l'interaction entre le concepteur et le système ressemble beaucoup à ce que fait un interviewer qui travaille avec ITAO lorsqu'il passe en revue un questionnaire. Le concepteur pourra peut-être avancer et reculer à volonté dans le questionnaire, choisissant les questions individuelles ou même des sections complètes avec leurs instructions "passez à" intactes.

Actuellement, ceux qui travaillent à la mise au point de systèmes pour enquêtes cherchent des moyens de représenter les liens qui existent entre les questions contenues dans leurs systèmes. Les concepteurs de systèmes ITAO, par exemple, doivent mettre au point des moyens pour suivre le cheminement d'un interviewé dans un questionnaire afin de vérifier si cette progression est conforme aux instructions "passez à" dudit questionnaire et de s'assurer qu'elle pourrait être reproduite si jamais il fallait revenir en arrière pour corriger une réponse antérieure (Nicholls et House, 1987 et Baker et Lefes, 1988). Les méthodes utilisées dans la plupart des systèmes ITAO, bien qu'elles soient

efficaces pour le genre d'opération que nous venons de décrire, ne semblent pas pouvoir très bien s'appliquer au problème tout à fait différent qu'est celui de la constitution de bases de données de questionnaires réalisées à partir de questionnaires multiples. On commence cependant à voir apparaître des applications très prometteuses de la théorie des graphes à la conception des questionnaires d'enquêtes. Le travail actuellement réalisé au Netherlands Central Bureau of Statistics (Willenborg, 1987) et à la UCLA (Futterman, 1987) offre des moyens de classer les questions d'après leur position dans des réseaux comme ceux qui sont créés par les instructions "passez à" de questionnaires complexes. Ce travail semble pouvoir réellement être appliqué à la conception de bases de données de questionnaires, bien que cela ne ce soit pas encore réalisé.

L'évaluation et le contrôle des questions

Lorsqu'on combine les questions et les réponses d'une enquête avec la distribution des réponses obtenues pour cette enquête, on obtient un outil puissant pour l'évaluation et le contrôle des questions. Une des principales préoccupations du concepteur de questionnaires est de savoir à quel point une question fonctionne. Les banques de questions donnent une idée aux concepteurs des modèles de réponses qu'ils sont susceptibles d'obtenir avec une structure de questions et un ensemble de réponses données. Quel taux de non-réponse une question produit-elle? Les réponses tendent-elles à se regrouper de façon à empêcher l'analyse? La formulation est-elle sujette à être mal interprétée par le répondant? Des questionnaires pilotes sont normalement utilisés pour étudier ces questions et améliorer le questionnaire avant le travail sur le terrain. Des banques de questions qui ont été bien mises au point, même si elles ne replacent pas les enquêtes pilotes, peuvent être utiles pour une première évaluation des questions et même pour effectuer des enquêtes pilotes plus efficaces et mieux orientées.

La mise sur pied de systèmes de saisie de données

La production de programmes informatiques pouvant commander des interviews automatisées est un autre emploi possible de la banque de questions. L'usage accru de technologies telles que l'ITAO et l'IPAO nous ont obligés à explorer des moyens de représenter les questionnaires sous des formes pouvant être utilisées par de tels systèmes. Pour la plupart des premiers systèmes ITAO, les concepteurs élaboraient des questionnaires en format imprimé habituel et le programmeur ou un autre technicien traduisait ensuite le questionnaire en un format lisible par machine que le système ITAO pouvait utiliser pour piloter une entrevue. Cette traduction se faisait normalement en un langage spécial de questionnaire, qui était ensuite traduit en un langage de programmation commun tel que FORTRAN. Ces systèmes représentaient ainsi le questionnaire comme un programme informatique. On sait que les utilisateurs de ces systèmes élaboraient des banques de questions rudimentaires en segmentant ces programmes et en conservant des blocs de code qui traitaient des questions précises ou des modules dans le questionnaire. Un tel code est ce que les programmeurs appellent un code réutilisable. Il peut permettre de gagner beaucoup de temps au cours de la réalisation d'enquêtes ultérieures utilisant les mêmes questions.

La tendance actuelle dans la conception de systèmes de saisie automatisée des données n'est pas orientée vers cette méthode mais plutôt vers la représentation des questions comme des éléments d'une base de données ou au moins comme un ensemble de tableaux représentant chacun un élément ou une question différente. Avec ces systèmes, le questionnaire est encore habituellement élaboré sur papier et introduit ensuite sur un écran d'ordinateur qui demande les différents composants de chaque question: le texte de la question, le type de réponse, les valeurs valides, le texte de réponse, les conditions booléennes permettant les sauts, et ainsi de suite. Ainsi, la première étape de la mise sur pied de l'ITAO et l'IPAO selon cette conception est de créer une banque de questions pour

chaque enquête. La prochaine étape logique serait donc de fusionner des bases de données de diverses enquêtes et de permettre aux concepteurs de choisir des questions ainsi que d'en ajouter.

Au fur et à mesure que progresse le travail réalisé avec les systèmes ITAO et IPAO utilisant des bases de données de questionnaires, nous nous approchons de plus en plus de l'utilisation de banques de données sophistiquées. Ces deux technologies exigent que nous explorions des moyens plus efficaces de représenter les questions et les questionnaires sous des formes lisibles par machine. Comme la production de questionnaires imprimés à partir d'éléments issus d'une banque de données, la production des applications des systèmes ITAO et IPAO requiert la mise au point de techniques permettant de représenter les liens qui existent entre les éléments, en particulier les instructions "passez à" du questionnaire.

La production de la documentation et du matériel de soutien

De bonnes données d'enquête requièrent une bonne documentation pour appuyer l'analyse. Tout dépendant des normes établies par chaque organisme d'enquête, cette documentation peut comprendre un document de codage détaillé comprenant le texte intégral des questions et des réponses ainsi que l'indication de la fréquence de chaque question, une structure d'enregistrement et de montages de progiciels d'analyse tels que ceux de SAS et de SPSS. Le questionnaire d'enquête constitue la base de cette production. La plupart des systèmes ITAO offrent déjà un de ces produits ou tous ces produits directement à partir de leur programme ou de leur fichier de questionnaires.

Les dictionnaires de données conçus pour commander l'extraction de variables par les analystes

À mesure que la technologie des bases de données évolue et que ses applications aux problèmes de la gestion des données des recherches par enquête sont élargies, on verra probablement apparaître une tendance vers les bases interactives de données d'enquêtes. La méthode actuelle qui consiste à créer de simples fichiers de données rectangulaires comportant des niveaux variés de documentation constitue un poids considérable pour les analystes. Les problèmes sont particulièrement aigus dans le cas d'enquêtes utilisant des questionnaires complexes et de multiples instruments, et de celles qui sont caractérisées par des conceptions longitudinales. On entend souvent parler d'analystes qui passent plus de temps à fusionner, à réduire et à manipuler des bases de données longues et complexes comportant plusieurs fichiers qu'à effectuer l'analyse elle-même. Le but de toute enquête devrait être de fournir une information fiable et pertinente sous une forme que les analystes peuvent utiliser. Pour la plupart des enquêtes réalisées aux États-Unis, ce but est plus un rêve qu'une réalité; on voit cependant apparaître de nouveaux développements très prometteurs (Robbin et David, 1988).

Les banques de questions pourraient être une partie de la solution au problème de la création de données prêtes à l'analyse. Imaginons qu'un analyste consulte une banque de questions avec l'intention de créer un fichier d'analyse à partir d'une enquête précise ou même un ensemble de fichiers à partir de plusieurs enquêtes. Une fois qu'il a choisi les questions, l'analyste peut soit effectuer directement l'extraction, soit produire un code de recouvrement pour un programme externe d'extraction exploité par la principale base de données de l'enquête. La banque de données est décrite ici comme une sorte de dictionnaire de données (métadonnées), qui définit et décrit tous les éléments d'une base de données relationnelle comprenant les véritables données de l'enquête. La banque de questions elle-même ne comprend pas les données d'enquête. Elle est plutôt un composant clé de tout système de données d'enquête qui répond aux besoins de tous les protagonistes de l'enquête: concepteurs, gestionnaires, analystes fonctionnels et chercheurs.

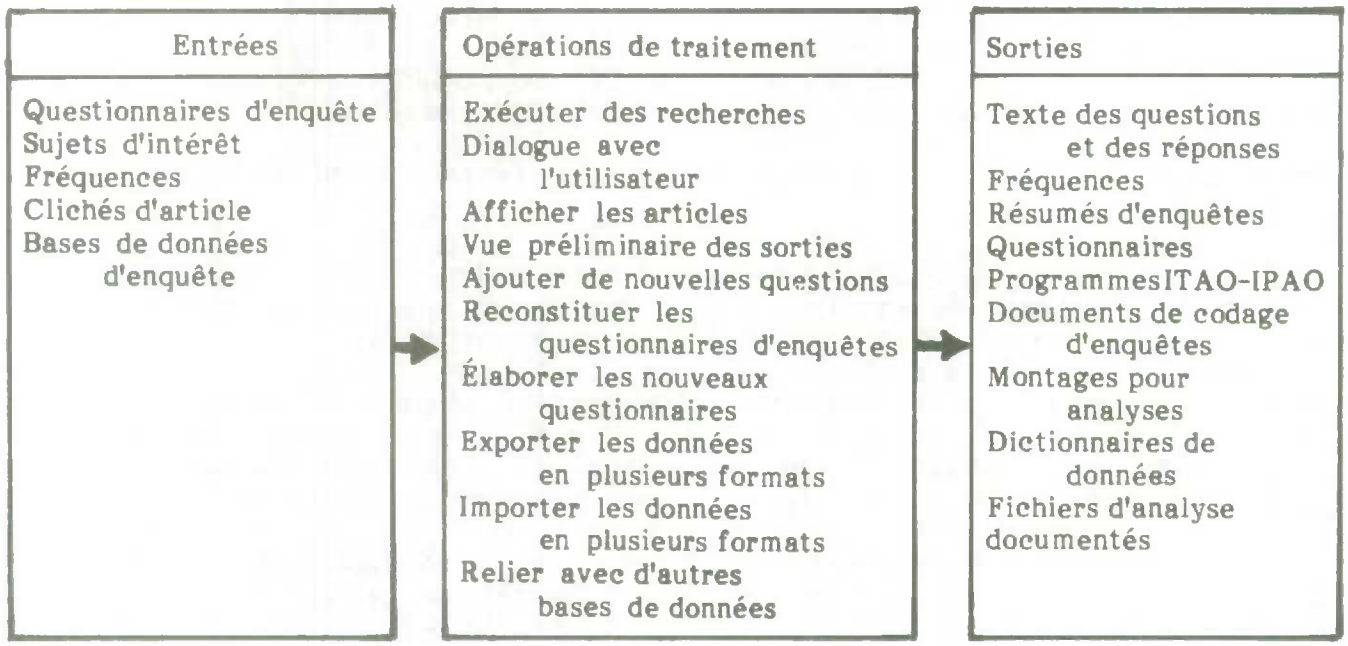
2. LA CONCEPTION DES BANQUES DE DONNÉES INFORMATISÉES

2.1 La configuration matérielle

La figure 1 est le modèle de la configuration matérielle très générale d'une banque de questions informatisée complète. La configuration matérielle est organisée en trois parties: les entrées du système, les opérations de traitement du système et les sorties du système. Les analystes fonctionnels élaborent normalement de tels modèles en sens inverse de la circulation des données, c'est-à-dire qu'ils commencent par les sorties et reculent jusqu'aux entrées en passant par les opérations de traitement. L'avantage principal de cette période de définition de la configuration matérielle est que l'élaboration d'un composant offre la plupart de l'information requise pour l'élaboration du suivant. Une fois qu'on connaît les sorties, il est plus facile de définir les opérations de traitement du système, et, une fois que les sorties et les opérations de traitement ont tous les deux été décrits, les entrées sont presque complètement définies.

Figure 1

Configuration matérielle



Les sorties indiquées dans la figure 1 sont essentiellement celles dont il a été question plus haut. Elles comprennent le texte intégral des questions et des réponses avec les fréquences, des résumés d'enquêtes (nom, organisme, taille de l'échantillon, etc.), les séries chronologiques des questions utilisées dans plusieurs enquêtes, les questionnaires imprimés, les codes ITAO-IPAO, les documents de codage des enquêtes (avec les fréquences), des montages tels que les commandes SAS et SPSS pour les programmes d'analyse, des dictionnaires de données pour les bases de données d'enquêtes, et même des fichiers d'analyse documentés.

Afin de produire ce large éventail de sorties, le système doit pouvoir réaliser un nombre d'opérations ou de traitements variés. Il doit pouvoir interagir avec les utilisateurs afin d'obtenir des critères pour la consultation et afficher les résultats. Il doit pouvoir recevoir de nouvelles questions dans sa base de données. Il doit pouvoir reconstituer des questionnaires entiers ou partiels à partir d'enquêtes

antérieures et produire de nouveaux questionnaires à partir de questions choisies par l'utilisateur. Le système doit pouvoir produire des sorties en formats variés - sur support papier et lisible par machine - et il doit pouvoir lire les données en formats variés. Finalement, afin de pouvoir créer des extractions, il doit pouvoir être relié à des bases de donées d'enquêtes.

Cet ensemble de sorties et d'opérations de traitement requiert cinq entrées: les questionnaires d'enquêtes, les sujets qui intéressent le concepteur ou l'analyste, les fréquences des éléments du questionnaire dans la base de données du questionnaire, les clichés d'article décrivant les bases de données d'enquête, et peut-être l'organisation des bases de données elles-mêmes.

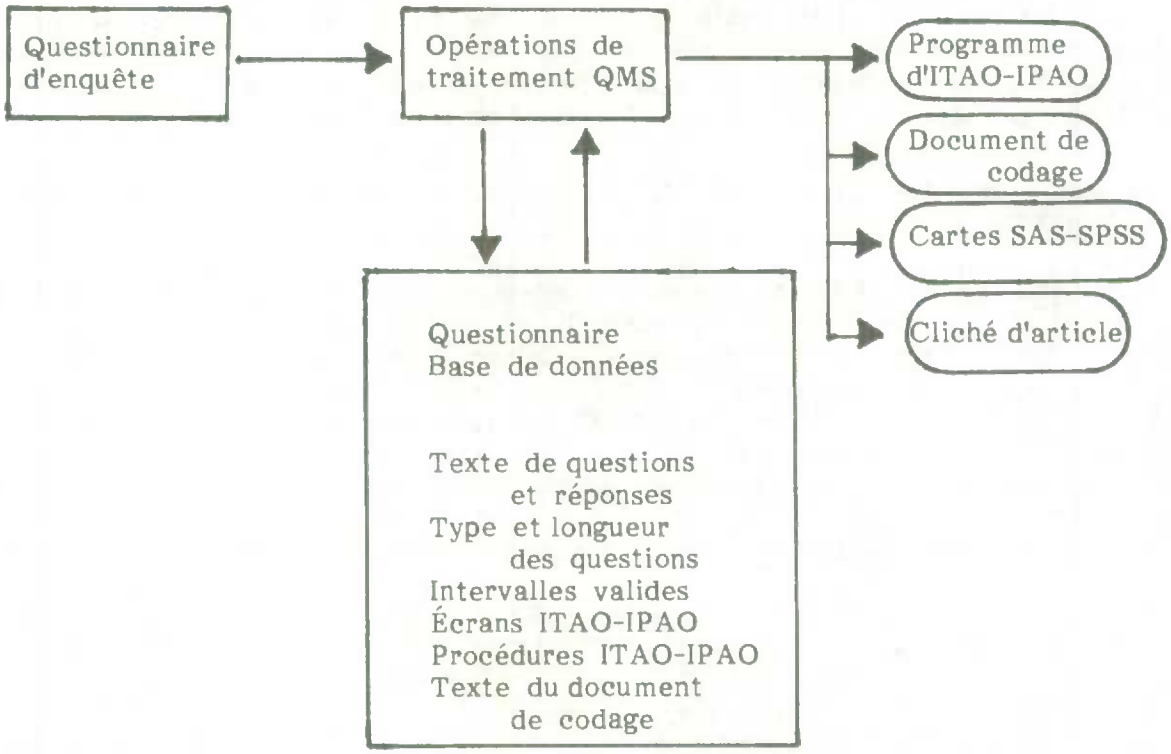
2.2 Le travail récemment réalisé au NORC

En 1985, le NORC entreprit un projet qui s'étendra sur plusieurs années visant une nouvelle conception de tous ses systèmes de production d'enquête (Baker, Beard et Taylor, 1986). Une des principales caractéristiques du nouveau système intégré est une base de données de questionnaire comprenant des programmes pouvant exécuter quelques-unes des fonctions d'une banque de données complète. Le composant de gestion de la base de données de questionnaire est appelé Questionnaire Management System (QMS).

QMS a été conçu d'après un modèle assez simple de données relationnelles. La figure 2 illustre la structure de base de son système. D'après sa conception actuelle, le système est utilisé exclusivement comme système d'alimentation pour les autres systèmes de production d'enquête. Son potentiel de mise au point de banques de données et de conception interactive de questionnaires n'a pas encore été exploité.

Figure 2

Le système de gestion de questionnaires du NORC



Le système est interactif, commandé par menu et fonctionne à base d'écrans à structure fixe. C'est un système à utilisateurs multiples appartenant à un vaste réseau local sur micro-ordinateurs (LAN) desservant tout le NORC. D'après la conception actuelle en phase de production, toutes les bases de données de questionnaires créées sont expressément conçues pour chaque enquête. Les concepteurs de questionnaires du NORC élaborent encore des questionnaires sur papier. Une fois qu'un questionnaire est terminé, le personnel chargé de la création et de l'entretien de la base de données QMS passent soigneusement en revue chacune des questions avec le concepteur. Un organigramme du questionnaire est également élaboré. Ayant en main le questionnaire imprimé, les questions révisées et l'organigramme comme références, l'utilisateur du QMS introduit sur les masques interactifs d'un écran d'OP le texte des questions et des réponses, le type et la longueur des questions, les noms variables des questions, les intervalles valides des réponses, et même de simples sauts. Des grilles sont également offertes pour d'autres formulations de questions telles que le texte du document de codage et le texte de l'ITAO et de l'IPAO. Le système offre un logiciel de coloriage pour l'élaboration à structure non imposée des écrans ITAO-IPAO et un éditeur pour l'écriture de procédures utilisées par le système ITAO-IPAO de NORC pour calculer les variables et les indices, pour effectuer des sauts complexes et pour contrôler tout mouvement non standard au cours d'interviews pilotées par ordinateur.

Pour les enquêtes longitudinales (le NORC en réalise plusieurs), les bases de données QMS d'une enquête servent de point de départ pour celles des suivis; il s'agit là d'une forme préliminaire de banque de questions. Notre expérience de modification de ces bases de données nous apprend beaucoup sur les limitations du système et sur la complexité qui sera nécessaire pour élever QMS au niveau d'une véritable banque de données.

À la fin de 1988, le NORC entreprendra la nouvelle conception du QMS dans l'intention d'en étendre les possibilités. Un de nos buts pour la conception est de rendre le système plus facile à utiliser afin que les concepteurs puissent l'utiliser directement pour monter leurs propres instruments. De plus, nous espérons pouvoir créer des bases de données de questionnaires à partir de plusieurs enquêtes, c'est-à-dire, créer une banque de questions. Nous voulons également améliorer la capacité du système d'importer des questionnaires qui sont déjà lisibles par machine afin qu'ils puissent être lus, analysés et introduits dans la base de données avec un minimum d'intervention de la part de l'utilisateur. Les sous-programmes qui produisent les codes pour ITAO-IPAO seront optimisés et le langage de procédure ITAO-IPAO sera plus facile d'utilisation.

3. L'AVENIR: LES DÉVELOPPEMENTS QUI NOUS MÈNENT VERS LES BANQUES DE QUESTIONS

La technologie informatique évolue rapidement, et chaque changement en élargit les applications. En même temps, la recherche et le développement dans le domaine des techniques d'enquête nous permettent de profiter mieux que jamais des bienfaits de la technologie informatique. Qu'on considère ces progrès individuellement ou globalement, il est évident qu'ils sont attribuables à une nette tendance de favoriser le développement des banques informatisées de données et d'améliorer leur conception. Cinq de ces développements, ou tendance, ressortent particulièrement.

3.1 L'utilisation croissante de l'ordinateur dans la recherche par enquête

Tel que le programme de ce congrès l'indique, la technologie informatique constitue le fondement d'une grande part de la recherche par enquête moderne. Toutes les étapes fondamentales du processus d'enquête - l'échantillonnage, la collecte, la préparation et l'analyse des données - sont gérées et réalisées avec des ordinateurs. La mise au point de systèmes de gestion de questionnaires et de saisie de données ont été particulièrement importantes.

Bien que l'ITAO n'ait encore que 15 ans, elle a déjà apporté des changements profonds à la façon dont nous menons les enquêtes. Elle est normalement moins coûteuse, et la qualité de ces données est supérieure à celle des méthodes traditionnelles (Groves et Kahn, 1979; Catlin et Ingram, 1988). On l'emploie de façon universelle dans les grands organismes d'enquête, et de plus en plus dans les petits organismes.

Une expérimentation très étendue a également été réalisée avec l'IPAO (Danielsson et Maarstad, 1982; Bemelmans-Spork et Sikkel, 1985; Rothschild et Wilson, 1988; Sebestik, Zelon, DeWitt, O'Reilly, et McGowan, 1988; et Foxon, 1988). Au début de la prochaine décennie, elle pourrait bien avoir remplacé le papier et le crayon dans les interviews en personnes menées par la plupart des organismes d'enquête du monde.

Étant donnée la tendance actuelle vers la saisie automatisée des données, les banques informatisées de questions semblent être l'outil idéal pour la conception et le gestion des questionnaires administrés par ITAO ou IPAO. En fait, la méthode selon laquelle les questionnaires sont conçus sur papier et envoyés au technicien spécialisé pour le montage de systèmes de saisie automatisée des données, se fait peu à peu remplacer par des systèmes de montage plus faciles à utiliser que le concepteur peut faire fonctionner directement. On mettra au point de nouveaux systèmes qui élaboreront la conception quelque peu primitive des systèmes actuels afin de créer un outil aux emplois multiples qui servira aux concepteurs de questionnaires et de systèmes et aux analystes d'enquêtes. Les banques de questions seront certainement une des principales caractéristiques de ces systèmes. L'expérience du NORC suggère que le premier but à viser est la création de banques maison, qui serviraient tout l'organisme et seraient peut-être grossièrement divisées par sujet.

3.2 Une analyse de questionnaire améliorée

Les banques informatisées de questions complètes requièrent de meilleures techniques pour la représentation des questions et des liens qui existent entre elles. Les techniques utilisées par la plupart des systèmes ITAO, bien qu'elles soient efficaces pour les applications propres à ces systèmes, ne peuvent être bien appliquées aux problèmes que posent les banques de questions. La recherche actuelle semble mettre l'accent sur les moyens de relier les questions selon l'ordre dans lequel elles sont posées. Les applications de la théorie des graphes à la représentation des instructions "passez à" des questionnaires, mentionnée plus haut dans cet exposé, sont extrêmement prometteuses, et nous attendons impatiemment la mise au point de systèmes qui incorporeront de telles conceptions.

Bien que la possibilité de coder des questions d'après la position qu'elles occupent dans la structure complexe des instructions "passez à" d'un questionnaire constitue un pas en avant, il faut également étudier des façons de regrouper les points du questionnaire selon l'information qu'ils contiennent. Les pratiques actuelles, comme la répartition des éléments de données par catégories de sujet, l'attribution de mots-clés et la recherche de questions par mots-clés dans leur contexte, paraissent presque primitives lorsqu'on les compare avec les techniques actuellement utilisées dans les entreprises pour la conception de systèmes d'information et de bases de données sophistiquées. La discipline de l'informatique et les techniques connexes telles que la modélisation des données (Martin, 1984) sont actuellement centrées sur la conception de bases de données relationnelles et de réseaux dans lesquels les éléments de données sont reliés selon leurs combinaisons naturelles pour constituer l'information. Appliquées aux problèmes du concepteur de bases de données, ces techniques pourraient aider à la conception de bases de données de questionnaire qui relient les questions non seulement selon l'ordre dans lequel elles sont posées aux répondants mais également selon l'information qu'elles contiennent.

La séquence qui suit illustre ces deux aspects du questionnaire.

Q1. Êtes-vous marié(e)?

1. Oui
2. Non (passez à Q5.)

Q2. Quel âge a votre conjoint?

Précisez: _____

Q3. A-t-il(elle) un emploi?

1. Oui
2. Non (passez à Q5.)

Q4. Quelle est sa profession?

Précisez: _____

Q5. Avez-vous des enfants qui vivent avec vous?

1. Oui
2. Non (passez à Q10.)

La figure 3 représente cette suite de questions telles que les voit l'analyste de questionnaire. Les questions apparaissent dans un ordre spécifique, chaque question définissant la prochaine question à poser. La figure 4 illustre la même suite de questions d'après la perspective de l'analyste des données. Les éléments de données sont reliés d'après les liens naturels qui existent entre eux. Là où l'analyste de questionnaire voit une série de noeuds et de dépendances dans un réseau, le chercheur voit une entité centrale (conjoint) ayant plusieurs attributs (âge, emploi, et profession). Là où l'analyste de questionnaire ne voit que des dépendances séquentielles, le chercheur voit des indépendances logiques. Autrement dit, d'après la perspective du questionnaire, l'âge précède la profession, et, donc, dans un sens l'âge définit la profession. Selon la perspective des données, l'âge et la profession ne sont liés que par leur appartenance à une entité commune appelée conjoint.

Figure 3

Configuration des questions

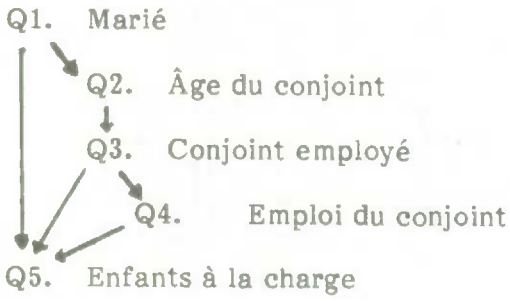
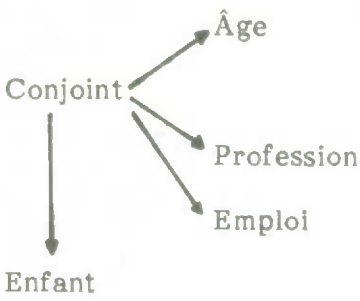


Figure 4

Configuration de l'information



3.3 Les systèmes experts

L'informatique et la conception de bases de données selon l'information contenue dans les éléments de données nous mènent vers le domaine de pointe que sont les systèmes experts. Un système expert est un système informatique conçu pour reproduire l'expertise humaine dans un domaine donné (Edmunds, 1988). Une caractéristique principale des systèmes experts est qu'ils sont articulés autour de la base de données, c'est-à-dire qu'ils utilisent des bases de données complexes appelées bases de connaissances plutôt que la logique complexe des programmes informatiques. On retrouve parmi les plus célèbres MYCIN, système expert au service des médecins pour diagnostiquer les maladies infectieuses, et HEARSAY, un système prototype capable de traduire le langage parlé, enregistré de vive voix.

Est-il possible de concevoir un système expert qui pourrait atteindre l'expertise du concepteur de questionnaires? Pour le moment, non. Plusieurs techniques de la conception de questionnaires pourraient certainement être intégrées à un système expert. Il s'agit de la sélection des questions candidates, de la structuration des instruments en modules, de l'établissement des instructions "passez à", et bien sûr, de la mise en page du questionnaire sur support papier ou pour un système de saisie de données tel que IPAO. Il pourrait même être possible à ce stade de mettre au point un système expert capable de faire un choix parmi des ensembles de catégories de réponses élaborés à partir des besoins d'information du concepteur. Bien que ces possibilités existent, la conception d'un système capable de rédiger des questions d'enquête ou même d'opérer un choix parmi différentes formulations de questions semble encore être un projet d'avenir. La technologie progressera cependant rapidement, de sorte que les concepteurs de systèmes d'enquête doivent donc commencer au plus tôt leurs essais afin que nous puissions bientôt profiter des avantages offerts par les systèmes experts.

3.4 Hypertexte

Hypertexte est le nom générique d'une nouvelle forme de document électronique. La caractéristique qui définit les documents hypertexte est leur dépendance envers l'information stockée dans un réseau de noeuds liés entre eux, par opposition à la structure linéaire des documents conventionnels (Smith et Weiss, 1988). Cette structure de réseau donne accès aux différentes parties du document par une méthode associative plutôt que par l'utilisation d'indexes et de tables des matières encombrants. Ainsi, hypertexte, ou hypermédia, comme on l'appelle parfois, organise et présente les données selon leur contenu et l'intérêt de l'utilisateur plutôt que selon leur disposition physique. La plupart des emplois actuels d'hypertexte entrent dans l'une des quatre catégories suivantes: documentation interactive pour les systèmes informatiques, ouvrages de référence dynamiques, publication automatisée de documents, ou aide pour l'enseignement de l'écriture (Barrett, 1987).

Il semble évident que hypertexte et ses structures conceptuelles peuvent être appliqués aux banques de questions. Par exemple, l'application actuelle de hypertexte à la mise en page sophistiquée de textes et à la constitution de documents pourrait être adoptée pour la production de questionnaires. La capacité de cette technologie de maintenir des bases de données de documents grâce à des réseaux complexes entre les éléments peut aider à la conception plus ou moins automatisée de questionnaires. Finalement, on se demande si les techniques de hypermédia permettraient la conception et l'administration via IPAO ou ITAO de questionnaires à "libre cours" qui suivront les propres associations de mémoire du répondant plutôt que la méthode traditionelle qui consiste à parcourir implacablement la route préconçue du questionnaire imprimé.

3.5 Les recherches en sciences de la connaissance

Un des phénomènes récents les plus significatifs dans le domaine des méthodes d'enquête est l'intérêt accru des chercheurs pour les théories de la psychologie de la connaissance (Jabine, Straf, Tanur et Tourangeau, 1984; Hippler, Schwarz et Sudman, 1987), et tout particulièrement pour leur application éventuelle au traitement de l'information et la conception de questionnaires. La notion qui soustend cet intérêt est la conviction qu'une meilleure compréhension de la façon dont nous traitons l'information et la consignons dans notre mémoire permettra la conception de questionnaires qui minimiseront les erreurs de réponse dues à des trous de mémoire chez le répondant.

Les travaux récents de Hastie suggèrent une façon d'appliquer aux problèmes du concepteur de questionnaires les résultats des recherches en sciences de la connaissance (Hastie, 1987). Il décrit les structures de la connaissance et de la mémoire en utilisant les techniques de représentation de la connaissance mises au point pour l'intelligence artificielle (IA). Ces techniques comprennent, par exemple, des modèles de réseaux associatifs où les concepts sémantiques sont représentés par des noeuds et les liens unissant ces concepts par des traits, un format semblable à celui illustré à la figure 4. Les modèles d'intelligence artificielle comprennent aussi un langage qui convient à la représentation des opérations sur des réseaux de bases de données, c'est-à-dire, les mêmes techniques que celles qu'utilise le concepteur de systèmes pour représenter les bases de connaissances pour les systèmes experts et les autres applications de l'IA. Ainsi, les sciences de la connaissance n'aident pas seulement à concevoir de meilleurs questionnaires mais aussi à trouver de meilleurs moyens de représenter l'information recueillie destinée à des programmes informatisés.

4. CONCLUSION

L'objectif de cet exposé était d'élargir vos connaissances sur les banques de questions et de vous informer sur les plus récentes recherches et applications innovatrices dans le domaine de la recherche par enquête. On y argumentait que les banques de questions ont un potentiel qui dépasse l'emploi qu'on en fait en tant qu'aide pour trouver des questions standard sur divers sujets d'intérêt. Les nouveaux emplois de la banque de questions comprenaient l'évaluation des questions, la production de questionnaires, soit sur support papier ou lisibles par machine, et la mise au point automatisée de métadonnées pour la création de dictionnaires de données et d'une documentation plus traditionnelle. Dans un avenir plus lointain, on s'attend à des bases de données de questionnaires qui modèleront une base de connaissances à être mise au point avec une enquête.

Actuellement, je ne connais aucune banque de questions qui correspond à la vaste définition exposée dans ce document. Néanmoins, les tendances dans la recherche par enquête et les développements de la technologie informatique nous mènent vers des systèmes présentant certaines de ces possibilités, sinon toutes. Dans la recherche par enquête, on assiste à l'automatisation continue de toutes les étapes du processus d'enquête. En même temps, les progrès de la technologie informatique nous éloignent des concepts de traitement des données qui jusqu'à maintenant ont guidé nos stratégies d'automatisation. L'avenir appartient au véritable traitement de l'information au moyen de bases de données relationnelles et de réseaux, à la représentation de la connaissance, et même aux systèmes experts.

Il existe probablement plusieurs façons de combler le fossé qui existe entre les pratiques actuelles de traitement dans la recherche par enquête et les nouveaux styles de traitement de l'information. On retrouve parmi les plus prometteuses les récentes recherches en sciences de la connaissance. Des études méthodologiques cherchant à améliorer les enquêtes par une conception claire de la façon dont les personnes traitent et

mettent en mémoire l'information dirigent la recherche par enquête vers l'intelligence artificielle comme moyen de comprendre la structure de la connaissance et de la mémoire et de représenter ces structures à l'ordinateur.

Une conception de la recherche par enquête comme étant de la cognitique est au coeur de toutes ces observations. Le but de toute enquête est la mise au point d'une base de données sur un domaine quelconque. Le concepteur de questionnaires travaille avec l'analyste et les cadres qui décident des politiques afin de déterminer l'information nécessaire à l'étude d'un problème; le concepteur établira ensuite un cadre de connaissances grâce auquel il organisera et présentera à l'analyste ou à l'administrateur l'information avec laquelle ceux-ci travailleront. Les enquêtes ont toujours été conçues de cette façon; le travail reste donc fondamentalement le même. Mais le cadre conceptuel et les outils utilisés pour ce travail changent, et continueront de changer de façon à rehausser l'efficacité et les possibilités de la conception, de la mise en oeuvre et de l'analyse des enquêtes.

Bibliographie

Baker, R.P., "Information Systems in Survey Research", Proceeding of the Bureau of the Census Third Annual Research Conference, Hyattsville (Md), 1987.

Baker, R.P., C. Beard et J. Taylor, "NISS: The NORC Integrated Survey System", Proceedings of the 1985 Public Health Conference on Records and Statistics, Hyattsville (Md), 12986.

Baker, R.P., et W.L. Lefes, "CATI System Design: A Review of Current Practice", sous la dir. de R.M. Groves et al, Telephone Survey Methodology, New York, John Wiley and Sons, à paraître.

Barrett, E., sous la dir. de Text, Context, and Hypertext, Cambridge (Ma), MIT Press, 1988.

Bemelmans-Spork, Mieke, E.J., et Dirk Sikkel, "Observation of Prices With Handheld Computers", Statistical Journal of the United Nations Economic Commission for Europe, 1985.

Catlin, G., et S. Ingram, "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Paper Methods in Centralized Interviewing", sous la dir. de R.M. Groves et al, Telephone Survey Methodology, New York, John Wiley and Sons, à paraître.

Edmunds, R.A., The Prentice Hall Guide to Expert Systems, Englewood Cliffs (NJ), Prentice Hall, 1987.

Foxon, J., "Field Trials in the Labour Force Survey Lap-Top Computer Project in November, 1987", Survey Methodology Bulletin, n° 23, p. 12-25.

Futterman, M., "An Approach to CATI Instrument Management Design Utilizing the Properties of Logical Structures", communication présentée lors de la International Conference on Telephone Survey Methodology, Charlotte (NC), novembre 1987.

Groves, R.M., et R.L., Khan, Surveys by Telephone: A National Comparison with Personal Interviews, New York, Academic Press, 1979.

Hastie, R., "Information Processing Theory for the Survey Researcher", sous la dir. de Hippler, H., N. Schwarz et S. Sudman, Social Information Processing and Survey Methodology, New York, Springer-Verlag, 1987.

Hippler, H., N. Schwarz et S. Sudman, sous la dir. de Social Information Processing and Survey Methodology, New York, Springer-Verlag, 1987.

Jabine, T., M. Straf, J. Tanur et R. Tourangeau, Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, Washington (DC), National Academy Press, 1984.

Martin, J., An Information Systems Manifesto, Englewood Cliffs (NJ), Prentice Hall, 1984.

Nicholls, W.L., et C.C. House, "Designing Questionnaires for Computer-Assisted Interviewing", Proceedings of the Bureau of the Census Third Annual Research Conference, Baltimore (MD), 1987.

Owen, K.V.N., et J.B. Mauro, The Question Finder, Richmond (Va), Smith & Flannery, Inc., 1988.

Robbin, A., et M. David, "SIPP ACCESS: Information Tools Improve Access to National Longitudinal Panel Surveys", RQ, n° 27 (4), p. 499-515.

Roper Center, User Manual for POLL: The Public Opinion Location Library, Storrs (Ct), Roper Center for Public Opinion Research, 1987.

Rothschild, Beth B., et Lucy B. Wilson, "A Nationwide Food Consumption Survey 1987: A Landmark Personal Interview Survey Using Laptop Computers", Proceeding of the Bureau of the Census Fourth Annual Research Conference, Arlington (Va), 1988.

Sebestik, Jutta, Harvey Zelon, Dale Dewitt, James M. O'Reilkly et Howard McGowan, "Initial Experiences with CAPI", Proceedings of the Bureau of the Census Fourth Annual Research Conference, Arlington (VA), 1988.

Smith, J.B., et S.F. Weiss, "Hypertext", Communications of the ACM, n° 31 (7), p. 816-819.

Willenborg, L.C.R.J., "The Routing Structure of Questionnaires", CBS Select 4: Automation in Survey Processing, Voorburg, Netherlands Central Bureau of Statistics, 1987.

LA COLLECTE ET LA SAISIE DES DONNÉES POUR LES ENQUÊTES

I. S. GREEN ET M. J. WADE,

Statistique Canada

COLLECTE ET SAISIE DES DONNÉES POUR LES ENQUETÊS

I. S. Green
M. J. Wade

Le 12 décembre 1988

Direction de l'informatique
Statistique Canada
Ottawa K1A 0T6

RÉSUMÉ

Les progrès récents en technique informatique nous suggèrent de nouvelles solutions pour régler les grands problèmes liés à la collecte et à la saisie des données, plus particulièrement dans un milieu où l'informatique est répartie. Nous étudions ici les possibilités qu'offrent ces solutions de rechange dans le cadre des récentes tentatives de Statistique Canada en vue de concevoir et de mettre au point un progiciel généralisé pour la collecte et la saisie des données.

INTRODUCTION

Dans de nombreux cas, la nouvelle technologie a déjà eu une incidence marquée sur la façon de diriger les activités du Bureau. La micro-informatique est devenue si courante et fait tellement partie de nos activités professionnelles quotidiennes qu'elle est maintenant passée au rang de seconde nature plutôt que de nouveauté qu'elle était il y a quelques années seulement. La nouvelle technologie a également rendu la mise en oeuvre réussie de systèmes complexes plus probable que ce n'était le cas dans les années 60 et 70, notre module de ***Généralisé de Vérification et d'Imputation*** en faisant foi.

Certaines de ces techniques ont évolué à un rythme surprenant; elles ont beaucoup modifié les relations de coût qui n'avaient pas bougé depuis des décennies. La technique de la micro-informatique a radicalement changé le rapport entre la puissance de calcul et les programmeurs : de nos jours, l'optimisation du temps de l'individu prédomine celle du temps d'ordinateur. Il s'agit là d'un renversement complet de la philosophie qui prévalait jusqu'à tout récemment.

L'apparition de nouveaux outils et de nouvelles techniques et le renversement spectaculaire des relations de coût, ajoutés aux changements internes d'importance instaurés par le Projet de remaniement des enquêtes-entreprises du Bureau, ont suscité un nouvel intérêt pour l'élaboration d'une méthode généralisée en vue du traitement des données d'enquête.

Cet intérêt a mené à la mise sur pied d'une équipe pour le ***Developpement de Fonctions Générales d'Enquête (DFGE)*** dans le cadre du projet de remaniement. L'equipe a pour mission de fournir une série de systèmes, fondés sur des méthodes normalisées, qui répondent aux exigences de la majorité des enquêtes-entreprises puis, éventuellement, des enquêtes sociales.

Du fait qu'elle consomme beaucoup de ressources et qu'elle exige une forte concentration en main-d'oeuvre, la fonction de collecte et de saisie des données a fait l'objet d'une attention accrue qui a mené à repenser en profondeur la manière dont se déroulent ces activités particulières. Par cette exploration, nous avons découvert quelques possibilités prometteuses.

Le mot édit dans ce texte peut-être defini de la façon suivante: un édit est une contrainte imposee sur les valeurs de variables données. Les éédits sont aussi utilisés pour déterminer les problèmes d'incohérence entre les réponses.

1. Le Milieu Actuel

Présentement, le cadre de collecte et de saisie des données de Statistique Canada reflète en grande partie les techniques qui ont caractérisé l'industrie du milieu des années 70 au début des années 80 et les facteurs monétaires connexes. Pendant cette période, tout comme au cours des années précédentes, le coût de la puissance de calcul était le facteur dominant pour déterminer l'emplacement, le type et l'étendue de l'automatisation. La tendance était à la centralisation des capacités de calcul et des ressources et le traitement se faisait beaucoup plus ***par lots*** que de façon ***interactive;*** nous ne pouvions tout simplement pas nous permettre de laisser un ordinateur inutilisé en attendant qu'un opérateur fasse l'introduction des données.

La technique des mini-ordinateurs nous a offert une solution de rechange beaucoup plus rentable que les ordinateurs centraux pour ce qui est de la conversion des données recueillies en format informatique et nous a permis une approche plus productive et versatile que les perforatrices à clavier. Au départ, quelques mini-ordinateurs ont été installés, un peu partout dans les services régionaux du Bureau, pour répondre surtout aux exigences de saisie des données de l'Enquête sur la population active par exemple. Bien que l'utilisation des mini-ordinateurs ait accru la productivité, ils ont surtout servi à déceler les erreurs de transcription d'entrer des données telles qui presentées. Cette méthode n'aidait aucunement l'interaction avec le répondant qui constitue la clé de la réduction des erreurs du répondant.

Cette automatisation restreinte a limité nos communications avec les répondants à ce qui suit :

- L'envoi et le retour par la poste d'un questionnaire normalisé.
- La collecte par téléphone et la transcription d'un questionnaire normalisé de suivi.
- Le report, à partir de documents de source, sur des formules de saisie de données.

Ce manque de versatilité a rendu la collecte et la saisie des données difficiles en introduisant des étapes intermédiaires entre la collecte réelle et la saisie. On a créé un environnement coûteux et vulnérable aux erreurs.

Dans une certaine mesure, ces étapes préliminaires ont également dicté ce qui se passe aux étapes ultérieures de la production pour obtenir des données traitées. Meme de nos jours, malgré les récentes initiatives entreprises au Bureau, le processus ultérieur typique se compose des étapes suivantes :

1. Les données recueillies sont transférées du mini-ordinateur à l'ordinateur principal.

2. Les édits par lot ont lieu et un *rapport aller-retour des échecs* est produit.

3. Les questionnaires sont extraits par les spécialistes.

4. Les corrections sont reportées sur les documents aller-retour et, s'il y a lieu, il y a suivi auprès du répondant.

5. Les mises à jour sont faites au moyen d'installations interactives de l'ordinateur principal ou d'un mini-ordinateur.

6. Selon l'endroit où les corrections sont faites, les étapes 1 à 5 ou 2 à 5 sont répétées jusqu'à ce qu'on manque de temps ou jusqu'à ce que cela cesse d'être rentable.

Toute cette approche se caractérise par beaucoup de transferts de données et d'interventions manuelles. Il est bien évident qu'au moment où l'on décèle les erreurs du répondant ainsi que les erreurs de transcription réintroduites, il est d'habitude trop tard pour tenter d'obtenir une réponse valide. Le seul recours pratique est l'imputation subséquente.

Au cours des dernières années, la nature de la collecte et de la saisie des données a changé, surtout du fait que le Bureau insiste davantage sur la régionalisation. Les versions interactives des édits traditionnelles de l'ordinateur central ont commencé à faire leur apparition sur mini-ordinateurs, ce qui fait passer l'accent de la simple détection des erreurs d'entrée à la tâche plus complexe de la détermination des erreurs du répondant. Bien que l'on ait retiré des avantages immédiats de cette approche, elle n'a pas été sans entraîner des coûts; les progiciels du secteur commercial servant à la saisie des données sont, en grande partie, incapables de répondre à de nombreux problèmes complexes. Conséquemment, les programmeurs ont créé d'importantes extensions sur mesure à ces progiciels et nous nous trouvons maintenant en charge d'un nombre sans cesse croissant d'applications propres aux enquêtes.

Parallèlement, l'équipement pour la saisie des données a Statistique Canada est devenu désuet et coûteux d'entretien; le système d'exploitation et les logiciels de saisie des données n'ont pas évolué. En bref, un remplacement s'impose maintenant. Essentiellement, deux options s'offrent à nous :

1. Moderniser le matériel actuel et transférer nos applications courantes à ce nouveau mode. De fait, nous serions à la merci d'un vendeur particulier, situation que nous connaissons bien!

2. Resister à la solution facile et tenter de trouver une approche moins axée vers les marques commerciales.

C'est cette dernière option que nous explorons.

2. Communication Avec le Répondant

Tout comme les autres organismes, nous tentons continuellement, à Statistique Canada, d'améliorer la qualité de nos données. En fin de compte, il faut pour cela un mécanisme de communication avec le répondant plus efficace. Notre stratégie est d'accorder plus de flexibilité au répondant et de communiquer avec lui d'une *fa çon plus personnelle.*

A la lumière des divers modes de collecte et de saisie des données qui sont devenus mieux reconnus dernièrement [4], nous avons décidé de laisser le répondant choisir la méthode qui lui convient. Pour les lecteurs qui ne connaissent pas bien ces modes d'opération, en voici de brèves descriptions.

- *INTERVIEW PAPIER ET CRAYON (IPC)*

  La méthode la plus connue et la mieux établie. Nous envoyons un questionnaire par la poste à une personne retenue pour faire partie de l'échantillon, en lui demandant de répondre au questionnaire et de le renvoyer par la poste une fois rempli; les données du questionnaire sont ensuite saisies.

- *INTERVIEW TÉLÉPHONIQUE ASSISTÉE PAR ORDINATEUR (ITAO)*

  Bien que cette technologie soit sur le marché depuis un certain temps déjà, son utilisation n'a pas été très répandue jusqu'à tout récemment alors que la baisse du coût du matériel l'a rendue plus accessible. Il s'agit d'une interview menée à un poste de travail interactif doté d'un appareil téléphonique. L'intervieweur lit au téléphone les questions qui sont affichees à l'écran, recueille les données et les entre directement dans le système. Bien que les systèmes d'interviews téléphoniques assistées par ordinateur soient essentiellement constitués ainsi, certaines applications très perfectionnées permettent également la maintenance de l'échantillon [5], l'ordonnancement des travaux et la production de divers rapports; le système d'interviews téléphoniques assistées par ordinateur du Bureau de recensement des États-Unis en est un excellent exemple.

- *INTERVIEW PERSONNELLE ASSISTÉE PAR ORDINATEUR (IPAO)*

  Micro-ordinateur portatif en main, l'intervieweur rend personnellement visite au répondant, à sa résidence ou à son lieu de travail. Avec cette méthode, l'échange d'information peut prendre la forme de questions et de réponses ou, si l'information se trouve sur des documents, les données peuvent être entrées directement; ainsi, la nécessité de transcrire les données en vue d'une saisie ultérieure est éliminée.

- *AUTO-INTERVIEW ASSISTÉE PAR ORDINATEUR (AIAO)*

  Avec cette méthode, l'organisme d'enquête fournit un logiciel de saisie des données au répondant; ce dernier peut utiliser son matériel, s'il est compatible, ou l'equipement fourni par l'organisme. Le répondant entre ses données, puis transfère ensuite les données de sortie à l'organisme.

- *SOURCES AUTOMATISEES*

  De nos jours, la grande majorité des entreprises et des organismes gouvernementaux sont automatisés. Par conséquent, bon nombre d'entre eux sont en mesure de fournir leurs données en format informatique. Statistique Canada profite de cette possibilité et reçoit une importante proportion des données dont le Bureau a besoin sur supports électroniques tels que les bandes magnétiques.

- *AUTRES*

  Il existe d'autres techniques pour la collecte et la saisie des données telles que les systèmes de traitement par reconnaissance vocale et de traitement d'image par ordinateur. Toutefois, à l'heure actuelle, ces techniques ne sont pas suffisamment évoluées pour être incorporées à un système généralisé de collecte et de saisie des données et, par conséquent, nous n'en parlerons pas davantage.

Afin de mieux répondre aux préférences personnelles des répondants parmi les divers modes de communication, nous pouvons également leur offrir un ***mode de communication mixte***. Par exemple, certains répondants peuvent souhaiter communiquer leurs réponses au cours d'une interview téléphonique assistée par ordinateur, mais insister néanmoins pour recevoir une copie papier du questionnaire. Dans d'autres cas, le répondant peut nous prévenir que si nous n'avons pas recu le questionnaire rempli à une certaine date, nous n'avons qu'a lui téléphoner; de fait, il s'agit d'une combinaison IPC et ITAO.

Même si nous avons tenté de répondre a toutes les préférences en offrant des modes de communication mixtes, les répondants font toujours face à trois problèmes persistants :

- *PROBLEME UN*

  Dans le cas des enquêtes qui visent des strates définies en termes généraux, par exemple Recensement des manufactures de Statistique Canada, nous risquons de nous retrouver avec un questionnaire très volumineux, dont de grandes parties ne s'appliquent pas à certains répondants.

- *PROBLEME DEUX*

  Certains répondants font partie des échantillons de plus d'une enquête et, en plus de faire l'objet de contacts accrus de la part du Bureau, ils peuvent être tenus de répondre (inutilement) à des questions communes à plusieurs enquêtes.

- *PROBLEME TROIS*

  Lorsque le questionnaire est volumineux ou varié, il se peut que l'entreprise répondante fasse appel à plus d'une personne pour y répondre. Ainsi, le questionnaire peut passer entre plusieurs mains avant d'être renvoyé par la poste.

Nous avons constaté que nous pouvons régler partiellement ces trois problèmes comme suit :

- en insistant davantage sur les caractéristiques propres à un répondant particulier comme base aux communications et

- en créant un modèle d'information sain, axé sur les données dont nous avons besoin plutôt que sur les aspects esthétiques tels que la structure et la phraséologie du questionnaire. En fin de compte, un modèle régi par les données plutôt que par les questions constitue une base plus cohésive pour l'élaboration des édits, la sélection des questions et l'élimination des contacts redondants et permet une plus grande ***accessibilité des données*** aux autres produits généralisés.

Pour ce qui est **du problème un,** supposons que ***l'enquête X*** exige des renseignements sur un produit donné et que nous sachions que ***le répondant X*** le fabrique, nous pouvons alors, à partir de ces renseignements, obtenir les metadonnées pertinentes relatives à cette marchandise. En ajoutant cela aux préférences connues du répondant quant au mode de collecte, nous pouvons choisir la méthode d'interview appropriée, (ex. ITAO plutôt que IPC). Inversement, si nous savons que le répondant X ***ne fabrique pas*** la marchandise en cause, les questions à ce sujet ***ne seront pas choisies.***

Il en résulte un instrument de collecte grandement simplifié et ***personnalisé,*** fondé avant tout sur les caractéristiques connues du répondant. Plus nous en savons sur le répondant, plus nous pouvons l'aider.

Pour ce qui est **du problème deux,** le fait d'être retenu pour plus d'une enquête est simplement une autre caractéristique connue du répondant X. Si nous adoptons ***une vue intégrée*** des besoins des enquêtes et que nous découvrons qu'elles possèdent des cycles complémentaires, ex. mensuellement, nous pouvons alors éliminer la redondance et produire un outil de collecte unique pour toutes les enquêtes en cause.

Pour trouver une solution **au problème trois,** il faut pousser un peu plus loin les solutions aux problèmes un et deux où un outil de collecte unique et réduit est produit au moyen de la sélection et de l'intégration des caractéristiques. Pour ce troisième problème, il faut créer quelques outils uniques. Il s'agit d'établir des caractéristiques d'un autre niveau. Si nous déterminons les exigences globales en données et que nous adoptons alors un point de vue logique, nous pouvons répartir ces caractéristiques en catégories (données comptables, données de production, données de ventes). Chaque catégorie peut être relevée au moyen de véhicules matériels ou électroniques de collecte distincts, destinés chacun à une ou plusieurs personnes dans une entreprise donnée.

On peut pousser plus loin l'utilisation des caractéristiques afin d'appuyer d'autres techniques qui réduisent le fardeau administratif du répondant et qui améliorent la qualité des données, par exemple les ***envois postaux échelonnés,*** plus particulièrement utiles dans le cas des enquêtes annuelles.

Il est bien évident que cette technique peut être portée à l'extrême. Le défi est donc de laisser les spécialistes et les méthodologistes determiner les limites pratiques de l'integration et de la caractérisation pour ce mécanisme de communication personnalisé.

3. Un Outil de Conception

Les nouvelles techniques ont certes ouvert la voie aux améliorations et aux changements phénoménaux dans la manière dont nous recueillons les données des répondants et les convertissons en formats informatiques. Au cours des dernières décennies, nous avons également connu des changements marquants dans la façon dont sont conçus les systèmes : chaque étape permet au concepteur de se concentrer davantage sur la solution du problème et de tenir moins compte des caractéristiques du matériel, de la structure des fichiers, des limites de procédures et des autres aspects de ce genre.

Nous retrouvons maintenant des techniques qui permettent à un concepteur d'aborder et de régler un problème donné par la méthode qui correspond le mieux à ses fins et à son style. En outre, ces techniques permettent au concepteur d'être plus conscient de l'ensemble des installations disponibles, ce qui encourage une meilleure utilisation. Tout comme c'est le cas avec les autres progrès en programmation, de moins en moins de compétences techniques sont nécessaires.

Au moyen des systèmes de programmation orientée-objet (POO), un environnement avec des fenetres à tâches multiples qui comprend des systèmes avec apparition de menus et utilization de souris, il y aura création d'interfaces qu'utiliseront les spécialistes et les méthodologistes pour informatiser leurs besoins en matière de collecte et de saisie des données pour les enquêtes, avec intervention minimale des préposés aux systèmes. Cela rapproche donc ces utilisateurs de la conception même du logiciel.

Grâce au *Système de Conception*, une personne pourra effectuer un grand nombre de tâches dans l'ordre logique voulu. Elle pourra, pour ne nommer que certaines des options, définir les exigences en données d'une enquête ou d'une série d'enquêtes particulières, décrire les règles d'édition qui serviront à filtrer les données et les questions utilisées pour les obtenir, et définir les règles qui feront le lien entre un certain répondant et ces données.

Cette approche permet d'économiser du temps, notamment dans trois domaines qui ont une importance particulière :

1. La conception du véhicule de collecte (questionnaire, scénario pour interview téléphonique assistée par ordinateur, etc.);

2. La specification des règles d'éditions;

3. La conception des affichages.

Une méthode assistée par ordinateur pour concevoir la présentation d'un questionnaire peut être beaucoup plus productive que les méthodes conventionnelles, même avec un système orienté-caractères, comme le démontrait le prototype généralisé mis au point à STC en 1987. Dans ce cas, la normalisation commençait au niveau des questions; ces dernières étaient grou-

pées en catégories logiques appelées blocs qui servaient à établir des questionnaires *personnalisés.*

Pour le texte d'une question, le concepteur pouvait aller jusqu'à trois lignes de 130 caractères chacune. Toutefois, le moniteur permettait de voir 80 caractères par ligne seulement; ainsi, si le texte dépassait cette limite, il fal lait avoir recours au déplacement horizontal pour le lire au complet. Il n'était pas possible de voir la question en entier si le texte dépassait cette limite.

L'impression (taille et type du jeu de caractères) des mots individuels d'une même question pouvait être modifiée par l'insertion de caractères parti culiers dans le texte qui servaient alors de signaux à d'imprimante.

Pour décrire un bloc (un groupe de questions avec un lien logique), on fournissait au concepteur un mécanisme lui permettant d'introduire le texte a l'appui sensiblement de la même manière que le texte d'une question. Après cela, des codes spéciaux étaient composés pour désigner les questions faisant partie du bloc, pour indiquer où laisser des blancs, pour préciser s'il y a lieu d'imprimer une ligne de démarcation, et ainsi de suite.

Pour déterminer de quoi une question, un bloc de questions ou un questionnaire auraient effectivement l'air, on retrouvait une fonction spéciale d'impression dans une autre partie du système afin de produire l'image sur imprimante à laser. Au moyen de cet outil de conception, il était possible de rédiger et d'imprimer un questionnaire au cours d'une séance de quelques heures seulement, ce qui constituait une amélioration marquée par rapport aux méthodes conventionnelles.

En utilisant un système de traitement multitâches, à base de graphiques et de fenêtres, combiné à un moniteur avec affichage large, nous pouvons obtenir les mêmes résultats que ci-dessus tout en réduisant *encore davantage le temps et l'effort requis.*

Avec des possibilités de graphiques et un grand écran, l'utilisateur peut com poser les questions directement à l'écran et les voir en entier, que les lignes dépassent 80 caractères de longueur ou non, puisque le concepteur choisie la dimension du texte apparaissant à l'écran. Le placement de caractères particuliers pour modifier la taille et le style du jeu d'impression pour les divers mots d'une question ne serait plus nécessaire étant donné que l'utilisation des clés de fonction (ou d'autres solutions novatrices) permettrait au concepteur de voir les jeux de caractères de façon dynamique - *une véritable vision du produit final (WYSIWYG).*

Si nous utilisons une programmation orientée-objet et que nous définissons les questionnaires, les affichages de saisie de données, les édits et la sécurité comme étant des objets, nous obtenons une très grande flexibilité pour la séance de conception,[1] plus particulièrement si le tout se déroule dans un milieu à tâches multiples. Par exemple si, après avoir déterminé une question, le concepteur souhaite poursuivre la conception du questionnaire et déterminer immédiatement une autre question de ce bloc, il peut le faire; il peut aussi passer à toute activité qui répond à ses fins pour le moment.

En plus de laisser au concepteur beaucoup de latitude quant à l'ordre à suivre pour ses activités, les nouvelles techniques de programmation permettent également la simulation de *la*

---

[1] En répartissant la conception en plusieurs éléments distincts, on devra tenir compte de l'integration de toutes les composantes nécessaires avant le transfert à la production.

façon de travailler au bureau. De fait, il s'agit de bureautique. Cela se précisera peut-être à mesure que nous poursuivons notre exemple de séance de conception du questionnaire.

Après avoir composé le texte de toutes les questions prévues dans un bloc (en supposant qu'il s'agit là de l'objet du concepteur), le travail peut-être conservé, l'image réduite au format de symbole graphique et déplacée à un autre endroit de l'écran, un peu comme si l'utilisateur plaçait un morceau de papier à un autre endroit sur son bureau.

Ensuite, le concepteur pourrait passer au mode de création de bloc et commencer à composer le texte à l'appui avec les installations utilisées auparavant pour la conception des questions. Après cela, au lieu d'indiquer au système au moyen de codes qu'il veut une ligne dans le questionnaire pour marquer la fin du bloc, il peut simplement *la dessiner*, un peu comme s'il utilisait une règle et un crayon.

Pour ramener les questions dans le bloc, plutôt que d'avoir à fournir une série de codes d'identité, le concepteur peut reprendre l'image réduite des questions qui se trouve à un autre endroit de l'écran, un peu comme s'il s'agissait de saisir une feuille de papier qui se trouve plus loin sur le bureau, l'ajuster a la taille voulue (ou utiliser des ciseaux pour couper le document à la taille voulue) et l'insérer dans l'image du bloc (coller les images de questions dans l'image de bloc). Il s'agit donc ***d'une opération de découpage et de montage graphique.***

Pour le moment, aucune impression du bloc n'est nécessaire; le concepteur peut voir exactement de quoi l'imprimé aura l'air en examinant ce qui se trouve à l'écran. Il peut, s'il le veut, demander l'impression, mais ***rien n'est encore définitif.***

Pour ce qui est de la présentation, les affichages pour la saisie des données ont la même conception que les questionnaires, sinon que ces derniers sont donnés sur papier alors que les premiers figurent à l'écran. Les deux contiennent des textes et des représentations graphiques; ainsi, un grand nombre d'installations et de techniques utilisées pour composer les questionnaires sont également des mécanismes permettant d'économiser du temps pour constituer les affichages.

Au cours de l'elaboration du prototype généralisé de collecte et de saisie des données, nous avons été en mesure d'établir de façon dynamique ***des affichages personnalisés implicites*** dans la mesure où nous definissions deux paramètres :

1. les données propres à une enquête;

2. le lien entre le répondant et les données (c.-à-d. la demande en données).

Au moyen de ces paramètres, le système de saisie pouvait dériver des messages-guides et déterminer où les placer à l'écran de sorte qu'ils soient utiles et respectent l'ordre dans lequel les questions sont posées au répondant. De plus, il pouvait faire des validations par défaut et générer de l'aide, selon le contexte, pour chacun des champs.

Voici où nous voulons en venir:

Si l'image de l'affichage de saisie des données n'est pas conceptuellement différente de celle du questionnaire et si nous possédons des renseignements sur les données d'enquête, y compris sur leur pertinence pour un répondant particulier, il s'ensuit que nous pouvons également générer implicitement et dynamiquement des questionnaires personnalisés.

Tout comme quand il s'agit d'établir les affichages de saisie de données, il y a des limites. Par exemple, le concepteur peut souhaiter apporter quelques modifications esthétiques à une image produite par le système. Le concept mérite toutefois une étude plus poussée.

Une troisième activité majeure permet de réduire considérablement l'effort et les compétences requises : la spécification de la validation. Les systèmes commerciaux ont d'excellentes approches à ce problème.

Bon nombre de fournisseurs de logiciels de saisie des données pour micro-ordinateur offrent, en plus de la possibilité d'établir l'affichage, des services de sélection et de spécification des règles de vérification qui rendent ces logiciels plus utiles que s'il s'agissait de simples substitutions à l'entrée des données par cartes ou bandes perforées.

Certains des vendeurs les plus populaires ont réglé le problème en offrant des bibliothèques de routines de validation généralisés et fournissent des services tels que l'aide contextuelle (messages affichés dans le cadre des interventions en cas de validations ratées). Les concepteurs de systèmes de saisie de données peuvent alors choisir pour leurs clients ces systèmes de validation parmi les routines prêtes à utiliser. Étant donné que ces routines vont de la séance de validation pilotée par paramètres, où il suffit de ***remplir les blancs***, à la séance de validation facultative, où il faut répondre par ***"oui" ou par "non"***, la nécessité de compétences particulières en programmation et de produits sur mesure a été grandement réduite.

Forcés d'améliorer leurs produits, certains vendeurs en ont poussé l'aspect fonctionnel en élaborant ***un langage de saisie des données*** que leurs clients peuvent rapidement apprendre et utiliser pour mettre au point des édits qui ne se trouvent pas dans les bibliothèques généralisées. Ces langages appuient l'exécution séquentielle, le branchement et l'itération et divers niveaux d'emboîtement; on peut y avoir accès à partir d'un grand nombre d'endroits dans une application donnée en tant que sous-programmes; de fait, il s'agit de langages de la troisième génération.

Bien qu'ils équivalent sous de nombreux rapports les extensions sur mesure, ces langages de saisie des données ont l'avantage d'être relativement faciles à apprendre et d'augmenter la versatilité d'un programme de saisie des données; le concepteur n'a pas à acheter de compilateur et d'éditeur de liens, ni à se familiariser avec eux.

Ces applications laissent toutefois à désirer dans deux domaines :

- Tout d'abord, ils sont nécessairement orientés vers l'industrie privée. Une forte proportion de leur marché se compose de détaillants, de banques et ainsi de suite. Par conséquent, ces langages visent à satisfaire la validation requise pour saisir les données d'un feuillet VISA, une tâche de moins grande envergure qu'un questionnaire de recensement.

- En deuxième lieu, nous pouvons raisonnablement supposer qu'en de nombreux cas les édits déterminées au moyen de langages de saisie des données seront très poussées. Bien que ces langages soient compilés, l'aide à l'auteur se limite surtout à des contrôles syntaxiques simples. On connaît des problèmes de logique au moment de l'exécution, problèmes pour lesquels on aura souvent besoin de l'aide d'un ***spécialiste en systèmes.***

Nous pouvons régler le premier problème en examinant de plus près certains de nos programmes de saisie faits sur mesure et de nos programmes d'édit de l'ordinateur central. Il est

fort probable qu'un grand nombre d'édits propres aux organismes de données statistiques peuvent être relevées, généralisées et ajoutées à la liste des produits disponibles sur le marché.

Pour ce qui est du second problème, il est possible d'acheter, ou de construire si nécessaire, des *analyseurs lexicaux* qui vérifient les règles d'édit d'un concepteur afin de déterminer si elles sont complètes, redondantes où contradictoires, ce qui réduit encore davantage la nécessité de compétences en programmation.

4. Le Milieu Futur

Statistique Canada présente des milieux informatiques variés. Des unités centrales en passant par les mini-ordinateurs et les micro-ordinateurs, pour aller jusqu'aux techniques hybrides, avec un mélange de réseaux, ce milieu s'étend sur des milliers de kilomètres et chevauche plusieurs fuseaux horaires : il s'agit d'une configuration véritablement répartie.

Bien que, dans ce cadre, les activités de collecte et de saisie des données puissent avoir lieu pratiquement n'importe où, se sont principalement deux unités organisationnelles particulières, la Division des opérations du bureau central (DOBC) et la Division des opérations des enquêtes (DOE), qui les font. Bien que la DOE dispose d'une représentation centrale, son cadre informatique est géographiquement réparti en huit bureaux régionaux et en services satellites connexes. La DOBC assure la responsabilité de collecte et de saisie des données pour les enquêtes qui se prêtent mieux à une approche centralisée. Dans les deux milieux, la mini-informatique domine.

Meme si nous nous attendons à ce que l'incidence de la nouvelle approche du Bureau à la collecte et à la saisie des données se répercute, ce sont ces deux unités organisationnelles, ainsi que leur clientèle respective, qui en retireront les plus grands bénéfices.

Une fois la mise en oeuvre terminée, nous prévoyons que deux cents enquêtes ou plus seront appuyées de cette façon, ce qui contribuera à créer un milieu très actif et très varié.

La meilleure façon de donner une idée concrète de ce que comporte la présente proposition est peut-être de décrire une enquête mensuelle hypothétique et d'en suivre un cycle complet. Avant de ce faire, nous devons poser les quatre hypothesès suivantes.

Hypothèse 1. L'enquête X est une enquête-entreprises et il existe une liste maîtresse ou une liste de l'ordinateur principal des entreprises canadiennes visées. A partir de cette liste, il est possible de retenir des échantillons et de déterminer les répondants de l'échantillon qui font partie de deux enquêtes ou plus.

Hypothèse 2. Grâce à un système de conception, les spécialistes et les méthodologistes ont décrit les exigences de l'enquête. Ces spécifications, sous forme de questions, d'affichages de saisie, de critères de sélection, de règles d'édition et de messages d'aide, ont été entièrement mises à l'essai grâce aux services de prototype du système de conception.

Hypothèse 3. L'enquête X a déjà été initialisée. Toutes les spécifications (règles) ont été transmises aux divers emplacements et chargées dans le système de production.

Hypothèse 4. La liste des entités de collecte des données pour l'enquête X est déterminée, chaque cycle, par un système externe à la collecte et à la saisie. Une entité de collecte peut se définir comme étant un enregistrement logique qui décrit les caractéristiques d'un répondant particulier, c'est-à-dire pour les enquêtes dont il fait partie, ses préférences de mode de communication, le nom et l'adresse des personnes contacts et les dates de suivi, et qui contient les renseignements précisant les données qui sont requises où qui ne le sont pas.

Pour amorcer le cycle, un système local de gestion de cas est initialisé. Chaque personne contact (il peut y en avoir plus d'une) au sein d'une même entreprise est traitée comme un cas distinct et reçoit son code propre d'identification.

Au cours de cette mise à jour, chaque cas est analysé.

Les caractéristiques des répondants qui préfèrent communiquer par IPC sont examinées. La demande en données est déterminée à partir d'une comparaison des caractéristiques individuelles de chaque cas aux critères de sélection. Au moyen de cette demande en données, on consulte ***une banque de questions normalisées;*** les questions et le texte complémentaire (tel que les instructions, les logos, les dates) sont choisis et le système génère, sur imprimante à laser, un questionnaire personnalisé, avec son propre code d'identification de cas, accompagné d'instructions pour l'expédition. Le système de gestion de cas est mis à jour en conséquence.[2]

Pour ce qui est des répondants qui préfèrent l'interview personnelle assistée par ordinateur (IPAO), les renseignements sur le cas sont acheminés à un intervieweur.

Lorsque l'intervieweur accepte la responsabilité des cas, ces derniers sont transférés sur une disquette qu'il pourra utiliser dans son micro-ordinateur portatif. De plus, les règles d'édition, les messages d'aide et les tables de référence liés à l'enquête X sont transférés a l'interpréteur de communication IPAO. Le système de gestion des cas est mis à jour pour tenir compte des attributions.

Dans le cas des répondants qui préfèrent effectuer eux-mêmes la saisie des données (AIAO) où les extraire de leur propre système informatique (sources automatisées), un rappel portant le nom, l'adresse et le code d'identification du cas est généré. Le système de gestion des cas est mis à jour pour indiquer l'impression et l'envoi du rappel, tout comme ce fut le cas pour les questionnaires.

Les autres répondants, qui ont choisi l'interview téléphonique assistée par ordinateur (ITAO) comme méthode de communication, sont placés en file d'attente en vue d'une interview. Tout comme pour les autres modes, le système de gestion des cas est mis à jour en conséquence.

Les contacts initiaux de tous les cas de l'enquête X ont alors été effectués ou amorcés. On s'intéresse maintenant moins à la gestion des renseignements qui quittent le lieu de travail qu'à la gestion des renseignements qui arrivent.

Pour les ITAO, les équipes d'intervieweurs sont désignées et les cas de la file d'attente sont attribuer à la personne désignée pour un cas donné ou à l'intervieweur qui est disponible. Pour une séance, les caractéristiques de chaque cas sont comparées aux critères de sélection, ce qui crée une demande de données sensiblement de la même manière que la production du ques-

---

[2] Etant donnée le grand nombre prévus de questionnaires, des imprimantes rapides seront utilisées.

tionnaire personnalisé. Dans ce cas, toutefois, au moment de la consultation de la banque de questions normalisées, le système utilise la version ITAO plutôt qu'IPC.

En même temps, les questionnaires remplis reviennent du terrain. Au fur et à mesure que ces questionnaires reviennent, les commis lisent les numéros d'identification de cas au moyen d'appareils de ROC et les renseignements sur le statut sont mis à jour. Simultanément, le système de gestion des cas attribue des groupes (lots) de questionnaires à des dossiers individuels qui sont ensuite attribués à des espaces d'entreposage.

A mesure que les opérateurs à la saisie des données de l'endroit deviennent disponibles, le gestionnaire leur remet les dossiers de l'enquête X. Avant que la saisie des données des questionnaires soit effectuée, les opérateurs font la lecture des numéros d'identification des dossiers et les inscrivent comme étant transférés à la saisie.

Une fois le dossier terminé, l'opérateur à la saisie des données signale le fait au système de contrôle des documents. Le système compare alors les questionnaires qui ont été entrés à ceux qui sont inscrits comme figurant dans le dossier et informe l'opérateur où le superviseur de tout écart. Une fois ces écarts réglés, la chemise est renvoyée à son espace de rangement et le système de contrôle des documents (du système de gestion des cas) est mis à jour.

Les questionnaires jugés incomplets où qui contiennent des erreurs des répondants sont alors ajoutés à la file d'attente en vue d'interviews de suivi assistées par ordinateur.

Les auto-interviews et les données de source automatisées, renvoyées par les services sur le terrain, sont automatiquement inscrites. Tous les cas de ce type font l'objet d'édition par lots. Ils sont ensuite traités de la même manière que s'il s'agissait d'un questionnaire en papier.

A mesure que les cas sont présentés aux fins de suivi aux lieux de travail, les intervieweurs peuvent :

- établir un ordre de priorité et obtenir les données manquantes dans un cas particulier où
- choisir uniquement les cas qui répondent a des critères de sélection prédéterminés (pondération de l'unité dans l'échantillon) ou
- suivre une règle d'observation minimale en tentant d'obtenir des réponses uniquement selon des critères de sélection prédéterminés, ex. l'importance relative de cette donnée dans l'ensemble du questionnaire où
- donner suite au cas selon une combinaison des diverses stratégies ci-dessus.

Essentiellement, cela couvre les principales composantes des activités à tout lieu de travail donné. Il faut cependant souligner que le système appuiera le transfert de cas *entre* lieux de travail.

Ce que le lecteur doit bien comprendre jusqu'ici :

- le niveau très élevé de l'automatisation et des communications en cause.

- la flexibilité du milieu : les modes mixtes, la génération de questionnaires personnalisés et de scénarios pour ITAO, etc.

- le fait que la gestion de cas peut assumer un rôle d'ensemble à l'égard de tous les modes de collecte et de saisie des données.

- le fait que la gestion des documents, avec la gestion de cas et la grande variété de mécanismes disponibles pour la collecte et la saisie des données, laisse au gestionnaire de l'emplacement une très grande latitude et beaucoup de contrôle.

- le fait que les éléments automatisés du système de production propose sont très intégrés.

5. Le Point de Vue du Gestionnaire

Le lecteur ne sera sans aucun doute pas étonné de savoir que les concepteurs de logiciels commerciaux pour la saisie des données visaient d'autres cibles que les organismes de données statistiques lorsqu'ils ont créés leurs produits. Dans de nombreux cas, ils cherchaient à maximiser la commerciabilité de leurs produits en s'adressant à une clientèle la plus générale possible. C'est à cause de ce phénomène que certaines des caractéristiques de ces produits ne nous sont pas très utiles, plus particulièrement les ***services d'information de gestion.***

En examinant les progiciels commerciaux pour mini-ordinateurs qui sont présentement utilisés au Bureau, ainsi que divers systèmes présentement disponibles orientés vers les micro-ordinateurs, nous avons constaté que les gestionnaires d'installations sont, dans l'ensemble, tenus par les vendeurs à une optique myope de frappes, de lots et de formes primitives de contrôle de la qualité. Pour régler ce problème, le Bureau a récemment conclu un marché en vue de l'élaboration d'un système de vérification par échantillons, à utiliser sur les mini-ordinateurs Honeywell, comme mode de rechange pour le contrôle de la qualité.

Le problème plus vaste demeure cependant.

Nous pouvons toutefois le réduire grandement par la mise en oeuvre d'un système véritablement généralisé de collecte et de saisie des données. Les gestionnaires, en intégrant une suite de modes pratiques et une approche globale à la gestion des cas, peuvent élargir leur vision des processus simplement en utilisant chaque mode comme jalon par rapport aux autres. Si nous arrivons à intégrer à la gestion des cas un système global de contrôle des documents, une plus grande partie du processus de production peut alors être contrôlée. Au-delà de ce point, il est possible à partir d'une base relationnelle de définir des normes et des indicateurs spéciaux. Ensemble, tous ces aspects permettront au gestionnaire d'avoir une beaucoup plus grande flexibilité pour plusieurs facettes de la surveillance et du contrôle exercés a l'égard des processus de collecte et de saisie des données.

Voici quelques exemples :

1. Si le modèle de données d'un système de production tient compte d'une combinaison de demandes d'enquête en un seul instrument de collecte, il va de soi que certaines formes d'analyses entre enquêtes peuvent avoir lieu au début de la production des données.

2. Si le gestionnaire d'installation ou d'enquête peut répartir les répondants en catégories - ceux qui répondent en temps opportun par IPC et ceux qui ne le font pas - il pourra alors faire passer les retardataires au mode ITAO. Cela permet d'accélérer le flot des données et accorde aux gestionnaires un meilleur contrôle sur l'utilisation des ressources.

3. Si nous considérons que le système de production est composé de "n" lieux de travail plutôt que d'avoir un système de production à chacun de ces lieux de travail, la direction peut alors prendre un certain recul face au processus de collecte et de saisie des données.

Les gestionnaires d'installations, les spécialistes et les méthodologistes font face à un défi difficile à relever : déterminer la meilleure manière d'utiliser ces possibilités prometteuses.

6. Perspectives d'Avenir

A l'heure actuelle, plus que jamais par le passé, la direction insiste sur les manières de réduire les coûts d'élaboration de logiciels et de soutien. Dans le cadre de notre stratégie à long terme, pour atteindre cet objectif :

- *NOUS RESPECTERONS DES NORMES DE TRANSFÉRABILITÉ*

  En utilisant des systèmes mis au point en langage de programmation "C" et en utilisant ce langage pour l'élaboration interne. Nous espérons que, au fur et à mesure que le matériel changera, la suite de logiciels généralisés pour la collecte et la saisie des données et les applications connexes pourront être transférées sans heurts, tout en évitant les coûts de conversion importants que nous avons connus par le passé.

- *NOUS UTILISERONS DAVANTAGE LES SYSTEMES GÉNÉRALISÉS*

  De cette manière, les spécialistes et les méthodologistes sont plus près de la conception véritable du logiciel, ce qui réduit la nécessité d'intermédiaires coûteux (spécialistes en systèmes). Étant donné qu'un moins grand nombre de gens doivent transmettre les exigences, nous pouvons obtenir un produit fini de meilleure qualité. De plus, les applications étant toutes en production sous l'égide d'un système de production, il suffit d'une seule équipe de systèmes pour la maintenance. Ces spécialistes peuvent alors offrir un soutien de plus haut calibre à leurs usagers.

- *NOUS RESPECTERONS LES NORMES INTERNATIONALES*

  En utilisant un logiciel de gestion des données (SGBDR) et des langages de programmation (langages rationnels SQL) reconnus et utilisés comme normes par la communaute statistique internationale, les occasions d'echange de logiciels et de compétences ainsi que d'autres formes de collaboration seront possibles, réduisant ainsi davantage le coût du logiciel.

Si l'on se tourne encore plus loin vers l'avenir, on peut voir que ces systèmes généralisés sont en réalité une application intelligente des progiciels qui, entre autres choses, permet l'interaction humaine pour la conception et la création des applications d'enquête. A ce titre, ils constituent une riche source des règles et des élément nécessaires pour créer les bases de connaissances que produiront la prochaine génération de logiciels de traitement d'enquête, à savoir les systèmes experts.

BIBLIOGRAPHY

1. Doucet, E. (1988) *Redefining Software Portability; A new Design Objective For Applications* Research And General Systems Subdivision, ISD, Informatics Branch, Statistics Canada.

2. Emery, D. (1987) *A Strategic Review Of Data Capture* **(Internal Document)** Research And General Systems Subdivision, ISD, Informatics Branch, Statistics Canada.

3. House, C.C. (1985) *Questionnaire Design With Computer-Assisted Telephone Interviewing* Journal Of Official Statistics, I(2). pp.209-219

4. Lyberg, H. (1985) *Plans For Computer-Assisted Data Collection At Statistics Sweden* Proceedings of the 45th Session. Bulletin of the International Statistical Institute, Book III, Topic 18.2, pp.1-11.

5. Nicholls, W. (1983) *Development Of CATI At The U.S. Census Bureau.* Proceedings of the American Statistical Association, Section on Survey Research Methods. Washington, D.C.

6. Shanks, J.M., Lavender, G.A., and Nichols, W.L. (1980), *Continuity And Change In Computer-Assisted Surveys - Development Of Berkeley SRC CATI.* Proceedings of the American Statistical Association, Section on Survey Research Methods, pp.507-512.

7. Von der Heyde, C.A. (1987) *Infrascope - The State Of The Art Of CATI In Germany.* Infratest Forschung, Stechprobenanlage Und Statistik, West Germany

LE TRAITEMENT D'ENQUÊTE ASSISTÉ PAR ORDINATEUR SUR LE PC ET LE PC-LAN

W. J. KELLER ET J. G. BETHLEHEM,

The Netherlands Central Bureau of Statistics

LE TRAITEMENT D'ENQUETE ASSISTÉ PAR ORDINATEUR SUR LE PC ET LE PC-LAN

Wouter J. Keller et Jelke G. Bethlehem
Netherlands Central Bureau of Statistics
Service d'automatisation
C.P. 959, 2270 AZ, Pays-Bas

RÉSUMÉ

Les centres nationaux de statistique utilisent toujours davantage l'ordinateur pour le traitement de l'information statistique. Vue l'énorme quantité de données à traiter, le micro-ordinateur ne semble pas être un instrument adéquat pour cette tâche. Le présent mémoire démontre le contraire: l'introduction de micro-ordinateurs a cependant un impact sur l'organisation des activités du service d'automatisation et des services chargées des sujets à traiter. Il sera donc nécessaire d'implanter une ferme politique de décentralisation du traitement des données et de normalisation des outils d'automatisation. On examine également la question des logiciels et du matériel, ainsi que l'utilisation du système Blaise pour le traitement d'enquêtes assisté par ordinateur.

1. LE PROCESSUS DE PRODUCTION DES STATISTIQUES

L'ordinateur a toujours joué un rôle important dans le traitement de l'information statistique. Déjà en 1880, Herman Hollerith construisait sa célèbre machine à cartes perforées, utilisée avec succès lors des recensements des États-Unis de 1880 et de 1890. Un des premiers ordinateurs fut fabriqué pour le U.S. Census Bureau. Au début, on n'utilisait l'ordinateur que pour le tri, le comptage, et la tabulation mais dans les années 1970, une nouvelle génération d'ordinateurs vit le jour qui permit aux statisticiens de réaliser des analyses statistiques grâce à des progiciels tels que SPSS, SAS, P-STAT et BMDP. Aujourd'hui, on utilise de plus en plus l'ordinateur pour réaliser les autres étapes du processus de production : l'édition des données, l'interview téléphonique assistée par ordinateur (CATI) et l'entrevue sur place assistée par ordinateur portatif (CAPI).

La production d'information statistique est un processus complexe. Les données sur les particuliers, les ménages et les commerces, recueillies au moyen d'enquêtes, doivent être trans-

formées en statistiques exactes. La première étape de ce processus est la collecte des données, habituellement réalisée au moyen d'un questionnaire. Le questionnaire assure que les mêmes questions soient posées à toutes les personnes interviewées mais ne permet pas d'éviter les erreurs. La deuxième étape est consacrée à l'édition des données, soit à la vérification des données et à l'élimination des erreurs dans la mesure du possible. Après l'édition, les données ne sont pas encore prêtes pour l'analyse et la tabulation. Il arrive presque toujours qu'elles ne soient pas représentatives de la population de laquelle elles sont issues. Ce problème est rectifié par le procédé nommé pondération, qui consiste à ajuster la valeur des éléments de l'enquête pour obtenir un ensemble de données plus correctes et représentatives. De là on passe à la dernière étape du processus de production, soit à la tabulation et à l'analyse. Ce mémoire porte sur le rôle de l'ordinateur dans la collecte, l'édition, la tabulation et l'analyse des données.

La mise en application des progrès récents de la technologie informatique est un prérequis pour la production efficace de statistiques de première qualité. Au Netherlands Central Bureau of Statistics (CBS), on utilise de plus en plus les micro-ordinateurs dans toutes les étapes du processus de production des statistiques. Le fait d'équiper un grand organisme avec de grandes quantités de micro-ordinateurs ouvre, d'une part, de nouvelles voies au traitement efficace de l'information statistique, mais crée aussi certains problèmes. Le CBS fait maintenant face au problème de la gestion d'un traitement décentralisé de l'information réalisé sur des centaines de micro-ordinateurs. La question du logiciel exige l'implantation d'une ferme politique de standardisation des outils d'automatisation. Ce mémoire porte une attention particulière au système Blaise de traitement d'enquête assisté par ordinateur, un outil qui réalise différentes étapes du processus de production des statistiques. Mais d'autres turelles pour constituer l'information. Appliquées aux problèmes du concepteur problèmes doivent également être réglés : la sécurité, la sauvegarde et l'archivage des données, les permis d'utilisation de logiciels et le partage des programmes et des fichiers de données. On propose la création de réseaux locaux et de grands réseaux comme solution à ces problèmes.

2. LA COLLECTE DES DONNÉES

Les organismes de production de statistiques recueillent la plupart de leurs données au moyen d'enquêtes par sondage. On compose à cet effet des questionnaires, ou des listes de questions à poser aux interviewés. Les questionnaires imprimés sont traditionnellement remplis au cours d'entrevues sur place: les enquêteurs rencontrent les interviewés, posent les questions et inscrivent les réponses obtenues sur les feuilles du questionnaire. La qualité des données ainsi recueillies est généralement satisfaisante. L'entrevue sur place est cependant coûteuse. Elle requiert un grand nombre d'interviewers, qui doivent tous beaucoup se déplacer. Voilà pourquoi on adopte volontiers l'entrevue téléphonique. L'enquêteur appelle l'interviewé depuis le bureau de statistique, évitant ainsi de se déplacer. L'entrevue téléphonique n'est cependant pas toujours réalisable: elle n'atteint que les personnes qui possèdent un téléphone et les questions ne doivent être ni trop nombreuses ni trop compliquées. Une entrevue par correspondance est encore plus économique: aucun enquêteur n'est nécessaire. Le questionnaire est expédié par la poste et on demande au destinataire de retourner la formule dûment remplie. L'ensoit possible d'envoyer des rappels, l'enquêteur n'étant pas en mesure d'exercer son pouvoir de persuasion, ce type d'enquête accuse un taux de participation plutôt faible.

Au cours de la dernière décennie, l'ordinateur a servi toujours plus à la collecte des données. Ce fut d'abord le cas pour l'entrevue par téléphone, alors que la CATI vit le jour. L'ordinateur établit l'horaire des appels et dirige l'entrevue. Il détermine quelles questions doivent être posées et vérifie la cohérence des réponses. Cette aide apportée par l'ordinateur a trois avantages majeurs. Premièrement, elle évite à l'enquêteur d'avoir à sélectionner la question suivante en fonction de la dernière réponse obtenue et éventuellement des instructions "passez à". Deuxièmement, cette technologie permet d'effectuer des vérifications pendant l'entrevue même. On peut ainsi discuter avec le sujet interrogé des erreurs décelées et apporter les corrections nécessaires, ce qui est exclu si la correction ne se fait que plus tard, au bureau de statistique. Troisièmement, les données étant déjà saisies au cours de l'entrevue, l'étape de saisie inférieure de données n'est plus nécessaire. La CATI est aujourd'hui une solide technique de collecte des données. (Pour une discussion complète des programmes CATI, voir, par exemple, Nichoes et Groves, 1986a, 1986b).

Depuis l'avènement récent du petit ordinateur portatif, l'enquêteur peut se rendre aux entrevues muni de son propre ordinateur. La CAPI réunit les avantages de l'entrevue sur place et de l'entrevue assistée par ordinateur. Les expériences réalisées par le Central Bureau of Statistics (CBS) avec le prédécesseur du système Blaise (voir Bemelmans-Spork et Sikkel, 1985a et 1985b), ont démontré que les enquêteurs peuvent travailler avec de petits ordinateurs et que l'emploi qu'ils en font au cours de l'entrevue n'entraîne aucune anxiété. Les résultats positifs de ces expériences emmenèrent le CBS à commencer l'utilisation généralisée de la CAPI dans le cadre d'une enquête réalisée en 1987, soit l'enquête sur la population active (voir Van Bastelaer et al, 1987a). Chaque mois, environ 300 enquêteurs, équipés de petits ordinateurs, ont visité 10 000 foyers. Après une journée consacrée aux entrevues, les piles de l'ordinateur portatif sont rechargées pendant la nuit, et l'information recueillie est transmise par téléphone au bureau central. Le lendemain matin, l'enquêteur retrouve l'appareil rechargé, avec un espace de travail entièrement libéré et prêt pour de nouvelles entrevues.

3. L'ÉDITION DES DONNÉES

Actuellement, la plupart des données sont recueillies de façon traditionnelle, à l'aide de formules imprimées et sans ordinateur. Les questionnaires ainsi remplis sont ensuite soumis à un traitement complet. La correction des erreurs est essentielle à la production de statistiques de qualité. On distingue trois types d'erreur. L'erreur d'intervalle se produit lorsqu'une donnée se situe à l'extérieur de l'ensemble de réponses valides possibles, par exemple, une personne âgée de 348 ans. Une erreur de cohérence se produit lorsqu'il y a un manque de cohésion entre les réponses recueillies pour un ensemble de questions. Les réponses "11 ans" pour l'âge et "marié" pour l'état civil sont valides, mais lorsque les deux ont été obtenues de la même personne, on fait définitivement face à un problème. Le troisième type d'erreur est l'erreur de routine ou de cheminement. Si l'enquêteur ou la personne interrogée ne suit pas la branche spécifiée ou les instructions "passez à", c'est-à-dire que si l'ordre du questionnaire n'est pas bien suivi, on aura répondu à des questions non pertinentes et des questions opportunes seront laissées sans réponse.

La correction des erreurs décelées s'impose, mais cette tâche peut être très difficile si elle est réalisée au bureau, une fois l'entrevue terminée. C'est particulièrement le cas pour les enquêtes réalisées auprès des ménages, car on ne peut plus communiquer avec les sujets interrogés; il faut alors trouver une autre façon de régler le problème. Il est parfois possible de faire

une approximation acceptable d'une valeur juste grâce à une technique d'imputation, mais il arrive plus souvent que la valeur inexacte soit remplacée par un code spécial indiquant que la valeur est "inconnue".

En plus de l'édition des données proprement dite, cette étape du processus de production comprend parfois une autre activité: le codage des questions ouvertes. Un exemple typique serait la question sur le travail de l'interviewé. Le traitement des questions est normalement très facile à réaliser lorsque l'interviewé choisit à partir d'une liste de réponses errogés; il faut alors trouver une autre façon de régler le problème. Il est parfois possible de faire une approximation acceptable d'une valeur juste grâce à une technique d'imputation, mais ipré-codées. Par contre, dans le cas de la question sur le travail de l'interviewé, cette liste de réponses précodées risque d'être très longue, et le sujet interrogé aura de la difficulté à sélectionner la bonne réponse. On peut éviter ce problème en permettant à l'interviewé de formuler sa propre réponse et en copiant littéralement celle-ci sur la formule. Les réponses doivent ensuite être classifiées pour qu'on puisse analyser l'information ainsi obtenue; la classification est une tâche qui prend du temps et qui ne peut être confiée qu'à des spécialistes. On explore actuellement des façons d'incorporer l'usage de l'ordinateur à ces activités. On a élaboré pour l'enquête sur la population active un système de codage assisté par ordinateur (voir Van Bastelaer et al, 1987b). Il s'agit bien d'un système "assisté par ordinateur" et non "automatisé", principalement parce qu'une grande partie du codage n'est pas indépendante du contexte; par exemple, le code assigné à un travail particulier dépend également du niveau d'éducation et du revenu. De plus, les techniques automatisées de reconnaissance de texte ne sont pas efficaces à 100 p. 100.

4. LA PONDÉRATION

Après l'édition des données, et éventuellement le codage des réponses ouvertes, on obtient un fichier "propre", c'est-à-dire sans erreur, mais qui n'est pas encore prêt pour l'analyse. D'abord, les échantillons choisis ont parfois des probabilités inégales, à savoir, les entreprises sélectionnées ont des probabilités proportionnelles à leur taille. En choisissant intelligemment les probabilités de sélection, on peut évaluer de façon plus exacte les paramètres de population, mais seulement en utilisant conjointement un procédé d'évaluation qui pourra corriger une telle inégalité. Ensuite, la représentativité est parfois affectée par la "non-réponse", c'est-à-dire, il arrive parfois qu'on n'obtienne pas de certaines unités statistiques, l'information nécessaire. Si les "non-répondants" ont un comportement qui diffère des caractéristiques de population faisant l'objet de l'enquête, les résultats obtenus seront faussés.

Afin de pallier aux effets de la non-réponse et de l'inégalité des probabilités de sélection, on utilise souvent la pondération. On attribue à chaque dossier des "poids" calculés de façon à ce que la distribution des caractéristiques telles que le sexe, l'âge, l'état civil et la région, qu'on retrouve dans l'échantillon, soit ajustée pour refléter la distribution réelle de ces caractéristiques dans l'ensemble de la population. Des recherches ont été effectuées au CBS afin d'améliorer les techniques de pondération (voir Bethlehem et Keller, 1987). La puissance accrue de l'ordinateur permet de mettre à exécution ces nouvelles techniques (voir Bethlehem, 1987).

5. LA TABULATION ET L'ANALYSE

Le fichier est enfin prêt pour l'analyse. La première étape de la phase d'analyse est presque toujours la tabulation des caractéristiques de base. A première vue, un tableau peut sembler très simple mais dans les faits il est tout autre. La composition des rangées et des colonnes (souvent constitués à partir de plus de variables qu'elles n'en affichent), l'affichage des quantités dans des cellules (comptes, moyennes et pourcentages), la façon dont les pourcentages sont calculés, le traitement des variables de plusieurs réponses, la position des totaux et des totaux partiels et beaucoup d'autres facteurs peuvent rendre pénible la constitution des tableaux. De bons progiciels de tabulation peuvent beaucoup aider; leur utilité dépend par contre de leurs possibilités, de la facilité d'utilisation de leur langage de commande et de leur capacité de produire une sortie prête à la photo.

Beaucoup d'organismes d'enquête font eux-mêmes l'analyse de leurs données afin d'en révéler les structures sous-jacentes et ainsi en retirer davantage d'information. L'information ainsi obtenue peut améliorer l'enquête qui suivra, rehaussant la qualité ou diminuant les coûts. Dans beaucoup de cas, les données sont nominales et il s'ensuit que des techniques reconnues d'analyse statistique pour les données d'intervalle ne peuvent être utilisées. Il existe bien sûr des techniques pour les données nominales, par exemple, l'analyse linéaire avec logs. Il y a toujours, cependant, un besoin pour d'autres techniques pouvant être utilisées pour de très grands ensembles de données. C'est ainsi que de nouvelles techniques, telles que ANOTA ou Correspondance Analysis, furent élaborées et mises en application. ANOTA (voir Keller et al, 1987) offre une autre possibilité pour l'analyse par logit qui étudie les rapports giciels de tabulation peuvent beaucoup aider; leur utilité dépend par contre de leurs possibilités, de la facilité d'utilisatioexistant entre les variables nominales dépendantes et un ensemble de variables explicatives. Correspondence Analysis (voir Bethlehem et Sikkel, 1987), est une technique étudiant de façon graphique les relations entre les variables nominales. Les deux techniques utilisent comme entrée des tables bivariées au lieu d'une matrice de données brutes, rendant même possible l'analyse à partir d'un grand ensemble de données sur micro-ordinateur.

6. LE SYSTÈME BLAISE

Blaise est le nom d'un nouveau système de traitement des données d'enquête exploité sur micro-ordinateur (ou sur un réseau de micro-ordinateurs) sous MS-DOS. Blaise effectue une grande partie du processus d'enquête: conception du questionnaire, collecte, édition et analyse des données. Le système Blaise est fondé sur un langage puissant et structuré (également appelé Blaise) qui décrit le questionnaire: les questions, les réponses possibles, la routine, le contrôle des intervalles et de la cohérence. Avec un questionnaire ainsi défini comme entrée, le système produit automatiquement des modules pour le traitement des données. D'abord, on peut produire un programme CADI (de saisie de données assistée par ordinateur) qui fournit un environnement intelligent et interactif pour la saisie et l'édition de données recueillies au moyen de questionnaires imprimés. Un autre module important est le programme CATI. Le système Blaise produit automatiquement le logiciel nécessaire à l'entrevue par téléphone pouvant, entre autre, gérer les appels. Un module qui ressemble au programme CATI est le programme CAPI qui est conçu pour l'interview sur place avec ordinateur portatif et qui peut également être produit par le système Blaise. De plus, Blaise produit automatiquement des montages et des fichiers de système pour le progiciel de statistique SPSS qui, dans un proche avenir, pourra être relié à d'autres progiciels du même type (par exemple PARADOX et STATA) grâce à des interfaces.

Les questionnaires traditionnels sont illisibles à cause des sauts imposés par les instructions "passez à". Le langage Blaise permet la conception structurée de questionnaires. La structure des questionnaires est donc très claire. On peut presque distinguer d'un seul coup d'oeil les questions à poser dans chaque cas. On trouve à la figure 1 un exemple d'un questionnaire Blaise simple.

Un questionnaire Blaise est comme une recette de tarte aux pommes, comprenant trois parties:

Première partie: la préparation des ingrédients

Pour préparer une tarte aux pommes, il faut d'abord que tous les ingrédients soient accessibles. Il en va de même pour un questionnaire: il faut d'abord définir toutes les questions, ce qu'on fait dans le "quest-paragraph". Une question doit contenir le nom de l'interviewé, le texte proprement dit de la question à poser et la définition des réponses valides. On distingue dans cet exemple trois types de réponses. "MarStat" et "Job" comportent des catégories de réponses définies par l'utilisateur. Il est à noter que de telles réponses sont constituées d'un court nom, utilisé pour le contrôle et pour suivre les instructions "passez à" dans d'autres parties du questionnaire, et d'une description complète utilisée comme explication dans les modules de logiciels produits. La question "Age" appelle une réponse en chiffres entiers limitée à l'intervalle 0,1,2,...,120. La question "KindJob" appelle une réponse ouverte, c'est-à-dire, un texte d'un maximum de 80 caractères.

Deuxième partie: mélanger les ingrédients

Au cours de la phase de préparation, on met ensemble les ingrédients dans le bon ordre. On procède de façon analogue dans le "route-paragraph" du questionnaire Blaise, qui indique selon chaque situation les questions auxquelles il faut répondre et dans quel ordre. La règle à suivre est simple: écrire le nom d'une question équivaut à poser cette question. Dans notre exemple, il est évident que les questions "Age" et "MarStat" sont posées à tous les répondants. On ne demande qu'aux personnes âgées de plus de 15 ans si elles ont un emploi. La question "KindJob" n'est posée qu'aux personnes âgées de plus de 15 ans qui ont un emploi.

Troisième partie: goûter le résultat

Après avoir terminé la tarte, on y goûte afin de savoir si elle est réussie. De même, un questionnaire doit comprendre des contrôles de cohérence. Dans un questionnaire, les contrôles sont précisés au "check-paragraph". On y détermine spécifiquement les contrôles de cohérence (les contrôles d'intervalle sont effectués automatiquement par le "quest-paragraph" et ceux de l'ordre des questions sont sous-entendus dans le "route-paragraph"). Le "check-paragraph" de notre exemple s'interprète comme suit: une personne âgée de moins de 15 ans ne peut être mariée.

La puissance du langage Blaise dépasse largement la description que nous venons d'en faire. Nous n'avons pas parlé du concept de bloc qui permet de bâtir des questionnaires longs et complexes à partir de sous-questionnaires. Nous n'avons pas mentionné le traitement des tables de questions sous forme de feuille de calcul, les variables locales, les types de réponses

```
QUESTIONNAIRE demo;
QUEST
     Seq (KEY)  "Numéro de séquence de
                l'interview": 1...10000;
     Age          "Quel âge a l'inter-
                viewé?": 0...120;
     MarStat      "Quel est l'état civil
                de l'interviewé?":
                  NotMar  "N'a jamais été
                          marié"
                  Marié   "Marié"
                  Divorce "Divorcé"
                  Widowed "Veuf ou veuve"
     Job          "L'interviewé a-t-il un
                emploi?":
                  (Yes "Oui, a un emploi",
                  Non  "Non, n'a pas d'
                  emploi");
     KindJob      "Quel type d'emploi
                l'interviewé a-t-il?":
                  STRING[80]

ROUTE
     Seq; Age; MarStat;
     IF Age > 15 THEN
          Job;
          IF Job = yes THEN
             KindJob;
          ENDIF;
     ENDIF;

CHECK
     IF Age < 15
       THEN MarStat = NotMar
     ENDIF;

ENDQUESTIONNAIRE.
```

Schéma 1: Un questionnaire BLAISE simple.

définies par l'utilisateur, l'utilisation de fichiers externes et de textes variables, pour ne nommer que quelques possibilités. On trouvera plus de détails sur ces sujets en consultant le manuel de référence sur le système Blaise (Bethlehem et al, 1987b).

Le système Blaise comprendra également un module pour le codage interactif, qui devrait pouvoir traiter la majorité des cas, laissant un petit nombre de cas difficiles au soin d'un spécialiste. Ce module comporte deux outils. Le premier met à exécution une méthode hiérarchique de codage. Le codage d'une réponse commence par l'entrée du premier chiffre du code en choisissant la catégorie appropriée à partir d'un menu. Après l'entrée d'un chiffre apparaît un menu subséquent contenant une sous-menu de la catégorie déjà sélectionnée. La description devient ainsi de plus en plus détaillée, jusqu'à l'obtention du dernier chiffre. Le deuxième outil effectue le codage selon une méthode alphabétique. Il tente de localiser une description parmi une liste alphabétique. Si la description est introuvable, la liste est affichée à partir du point le plus proche de la description qui a été introduite. La liste peut être construite de façon à ce que presque n'importe quelle description y soit présente, y compris les permutations.

Cette méthode est avantageuse parce qu'elle est simple, rapide et maîtrisable. Les deux outils de codage peuvent être utilisés simultanément.

Une application importante du traitement des données que le système Blaise peut produire est la machine CADI: un système intelligent et interactif servant à la saisie et à l'édition des données. Les données peuvent être traitées de deux façons. La première est la suivante: le spécialiste du sujet à l'étude s'assoit à un micro-ordinateur avec une pile de formules qu'il traite l'une après l'autre. Il introduit les données au bon endroit et, à la fin de chaque formule, il active l'option de contrôle pour vérifier si l'ordre des questions a été respecté et s'il n'y a pas d'incohérence (les erreurs d'intervalle ayant été décelées lors de la saisie des données). Les erreurs décelées sont mises en évidence et expliquées à l'écran. Elles peuvent être corrigées en consultant la formule ou en téléphonant à la personne qui a fourni l'information. Après l'élimination des erreurs, un dossier "propre" est versé au fichier. Si le spécialiste n'arrive pas à produire un dossier sans erreur, il peut le mettre dans un fichier à part réservé aux dossiers à problème. Ces dossiers seront traités ultérieurement, également avec l'aide de la machine CADI. La seconde façon de traiter les données avec la machine CADI est la suivante: les données sont introduites par des agents de saisie des données, sans vérification. Les dossiers du fichier peuvent être édités par la machine CADI. On les extrait du fichier un par un, on active l'option de contrôle et les erreurs ainsi décelées sont ensuite examinées et corrigées.

La machine CADI sert surtout à traiter les données ayant déjà été recueillies sur des formules imprimées. Le système Blaise peut cependant être utilisé pour la collecte des données également. Il peut produire les logiciels nécessaires pour réaliser des entrevues par CATI et par CAPI. Un programme CATI suit un cheminement dynamique. Après la saisie d'une réponse, la machine CATI décide quelle question doit suivre et passe automatiquement à celle-ci. Notez ici la différence avec la machine CADI, qui suit un cheminement statique. Cette différence est due à l'usage tout à fait distinct qu'on peut faire des deux machines. Avec la CADI, on cherche à copier les données d'une formule et à en déceler les erreurs; avec la CATI, on cherche à éviter toute forme d'erreur.

La machine CATI décèle les erreurs de façon dynamique. Dès qu'une erreur de cohérence est trouvée, un message d'erreur est affiché à l'écran, ainsi que la liste des questions reliées. Via un menu, l'utilisateur peut choisir la question à laquelle il doit passer pour corriger l'erreur.

Le système Blaise peut également produire un programme CAPI, qui peut être utilisé avec un ordinateur portatif. La CAPI offre les avantages de l'interview sur place en plus des qualités de la CATI. Le programme pour ordinateur portatif mène l'entrevue: il en détermine la routine (le cheminement) et il vérifie la cohérence des réponses. Puisque le contrôle se fait pendant l'entrevue, les erreurs peuvent être corrigées sur-le-champ. Après une entrevue réussie, le dossier sans erreur est mis en mémoire dans l'ordinateur portatif. Les données sont transmises au bureau plus tard au moyen d'un modem et d'un téléphone, ou expédiées sur disquette.

7. LA GESTION DE 1 000 MICRO-ORDINATEURS

Un micro-ordinateur est un précieux instrument de traitement des données statistiques; il faut cependant en distribuer un grand nombre à l'intérieur de l'organisme de statistique pour que son utilisation soit efficace. S'il s'agit de micro-ordinateurs autonomes, des problèmes se posent inévitablement. Le premier problème est celui de la sécurité. Les données statistiques se fondent dur des renseignements personnels qui devraient normalement rester confidentielles. On ne peut guère garantir le secret si les données sont stockées sur des disques rigides de micro-ordinateur. Un deuxième problème est la sauvegarde et l'archivage des données, qui se trouvent sur des disquettes et des disques rigides un peu partout dans un grand édifice. Les permis d'utilisation des logiciels peuvent également créer des difficultés: pour beaucoup de micro-ordinateurs, il faut un grand nombre de progiciels, et donc un grand nombre de permis. Le contrôle de la version n'est pas facile non plus. Enfin, le partage des programmes, des fichiers de données et du matériel ne peut se faire efficacement.

Le CBS a résolu ces problèmes en constituant des réseaux. Chaque service a son propre réseau local (LAN), exploité avec des logiciels Novell, et du matériel de réseau Datapoint. De 10 à 40 micro-ordinateurs sont branchés à un serveur de fichiers d'une capacité de 150 à 300 Mb. La sécurité est assurée par l'utilisation d'un mot de passe pour la procédure d'entrée en communication, par le chiffrement et par une encoche de protection d'écriture sur les disquettes. On prévoit le jour où les disquettes mêmes seront superflues, ce qui entraînera la suppression des unités de disquette. Comme tous les logiciels et les fichiers de données sont emmagasinés sur le même serveur de fichier, l'utilisation en partage devient très simple.

Au CBS, il y a environ 50 réseaux locaux. A chaque réseau est assigné un administrateur et un administrateur adjoint. Ces deux personnes jouent un rôle essentiel dans l'installation et dans l'entretien des logiciels, ainsi que dans l'assistance à l'utilisateur. Tous les réseaux locaux sont reliés à un réseau Datapoint de base (qui sera bientôt remplacé par Ethernet). La grande unité centrale CDC Cyber et les mini-ordinateurs CDC 930 sont branchés au grand réseau, grâce à Ethernet et à des passerelles. Les deux emplacements du CBS (Voorburg et Heerlen, à 300 km l'un de l'autre) sont reliés par un lien de 2 Mbps.

L'archivage et la sauvegarde sont réalisés de manière centralisée par le service d'automatisation (sauvegarde totale de 10 giga-octets!). Il est évident que le contrôle de la version et la mise à jour du logiciel sont simples à réaliser dans un tel cadre. Tous les permis d'utilisation des logiciels sont basés sur l'utilisation simultanée, contrôlée par des logiciels maison.

Le conseil d'administration fait procéder deux fois par an à l'allocation des micro-ordinateurs aux services, par lots de 200. Le nombre de permis d'utilisation de logiciels (exprimé en "calculs" simultanés) et d'ensembles de documentation, la quantité d'espace disques et le nombre d'imprimantes sont déterminés par de simples règles fondées sur le nombre de micro-ordinateurs alloués à chaque service. Le besoin de former un grand nombre d'utilisateurs d'OP fait que le CBS dispense en moyenne 50 cours d'une journée chacun par mois (utilisant trois salles de cours bien équipées tous les jours ouvrables).

8. CONCLUSION

L'avènement du micro-ordinateur a eu un impact considérable sur le travail des bureaux nationaux de statistique. A mesure que croît le nombre des statisticiens inexpérimentés qui doivent s'en servir pour accomplir leur travail, le besoin d'un logiciel tel que le système Blaise ne cessera d'augmenter le besoin. La puissance de ce système provient, d'une part, de la cohérence qu'il assure à chaque étape de collecte et de traitement des données; tout le processus en devient plus facile à gérer et à maîtriser. Il assure d'autre part la normalisation entre les services. Comme ceux-ci utilisent tous les mêmes logiciels pour effectuer leurs enquêtes, l'échange d'information est plus aisé et moins sujet à erreur.

Actuellement, l'ordinateur est parfaitement intégré au processus de production de statistiques. Le CBS est équipé de centaines de micro-ordinateurs. Chaque service a son propre réseau local, et tous les réseaux locaux sont reliés entre eux et reliés avec l'unité centrale et les minis du service d'automatisation. Un traitement simple des données électroniques est maintenant effectué par les services mêmes qui sont chargés du sujet à l'étude; la conception et l'entretien de systèmes d'information complexes sont laissés aux spécialistes en informatique du service d'automatisation. Tout travail interactif (comprenant la saisie et l'édition des données) est réalisé sur micro-ordinateur, alors que le traitement par lots (la pondération et la tabulation) se fait sur l'unité centrale et les mini-ordinateurs.

L'utilisation intensive et plus sophistiquée des ordinateurs améliore la qualité des données et l'efficacité, mais crée aussi de nouveaux problèmes. La décentralisation des activités informatiques a fait naître le besoin de normaliser les outils de d'automatisation (un seul système de traitement de texte, un seul progiciel pour les bases de données, un seul tableur, un seul progiciel d'analyse, etc.). Dans un tel milieu, la formation et le concours d'experts sont essentiels, ainsi que la coordination. En d'autres termes, l'emploi décentralisé dans un milieu centralisé.

BIBLIOGRAPHIE

1. Van Bastelaer, A.M.L., F.A.M. Kerssemakers et D.Sikkel (1987a), "A test of the Continuous Labour Force Survey with hand-held computers : Interviewer behaviour and data quality", CBS select 4, Automation in Survey Processing, Voorburg, Netherlands Central Bureau of Statistics.

2. Van Bastelaer, A.M.L., L.M.B.P. Hofman et J.K. Jonker (1987b), "Computer Assisted Coding of Occupation", CBS select 4, Automation in Survey Processing, Voorburg, Netherlands Central Bureau of Statistics.

3. Bemelmans-Spork, E.J., et D. Sikkel (1985a), "Observation of prices with hand-held computers", Statistical Journal of the United Nations Economic Commission of Europe, vol. 3, no 2.

4. Bemelmans-Spork, E.J., et D. Sikkel (1985b), "Data Collection with hand-held computers", Proceedings of the 45th session, International Statistical Institute, Tome III, sujet 18.3.

5. Bethlehem, J.G. (1985), LINWEIGHT User Manual, CBS report, Voorburg, Netherlands Central Bureau of Statistics.

6. Bethlehem, J.G., D. Denteneer, A.J. Hundepool, et W.J. Keller (1987a), BLAISE 1.1 / A first acquaintance, CBS report, Voorburg, Netherlands Central Bureau of Statistics. Bethlehem, J.G., D. Denteneer, A.J. Hundepool, et W.J. Keller (1987b), BLAISE 1.1 / Language reference manual, CBS report, Voorburg, Netherlands Central Bureau of Statistics.

7. Bethlehem, J.G., D. Denteneer, A.J. Hundepool, et W.J. Keller (1987c), BLAISE 1.1 / The data entry machine, CBS report, Voorburg, Netherlands Central Bureau of Statistics.

8. Bethlehem, J.G., D. Denteneer, A.J. Hundepool, W.J. Keller, et M.H. Schuerhoff (1987d), BLAISE 1.1 / System description, CBS report, Voorburg, Netherlands Central Bureau of Statistics.

9. Bethlehem, J.G. et W.J. Keller (1987), "Linear Weighting of Sample Survey Data", Journal of Official Statistics, Vol. 3, p. 141-154.

10. Bethlehem, J.G. et D. Sikkel (1985), CORAN User Manual, CBS report, Voorburg, Netherlands Central Bureau of Statistics.

11. Keller, W.J., A. Verbeek et J.G. Bethlehem (1985), ANOTA : Analysis of Tables, CBS report, Voorburg, Netherlands Central Bureau of Statistics.

12. Nicholls II, W.L., et R.M. Groves (1986a), "The status of computer telephone interviewing : Part I- Introduction and impact and cost and timeliness of survey data", Journal of Official Statistics, no 2, p. 93-115. Nicholls II, W.L., et R.M. Groves (1986b), "The status of computer assisted interviewing : Part II- Data quality issues", Journal of Official Statistics, no 2, p. 117-134.

L'UTILISATION DES MICRO-ORDINATEURS POUR LE TRAITEMENT DES DONNÉES DE RECENSEMENT DANS DES PAYS EN VOIE DE DÉVELOPPEMENT: UNE MISE À JOUR

V. TORO ET K. CHAMBERLAIN,

US Bureau of The Census

Introduction

Ce document constitue une mise à jour du document de Vivian Toro et Thomas Melaney, intitulé "L'utilisation des micro-ordinateurs pour le traitement des données de recensement dans des pays en voie de développement", et présenté dans le cadre des réunions de l'American Statistical Association, en août 1987. Dans ce document, les auteurs examinent les solutions qu'offrent les micro-ordinateurs aux nombreux problèmes éprouvés dans les pays en voie de développement lors du traitement des données de recensement au moyen d'ordinateurs principaux. Les auteurs y décrivent également le Système de Traitement Intégré sur Micro-ordinateurs (STIM) élaboré par l'International Statistical Programs Center (ISPC), du U.S. Bureau of the Census, et présentent trois études de cas sur l'utilisation du STIM par les bureaux de statistiques des pays en voie de développement.

Les expériences positives d'un certain nombre de pays en voie de développement confirment que les micro-ordinateurs sont des outils technologiques sûrs et rentables pour le traitement des données de recensement. Les progrès technologiques constants et le perfectionnement des logiciels pour micro-ordinateurs rendent leur utilisation encore plus avantageuse.

Ce document porte principalement sur les récentes expériences du Burkina Faso et du Sénégal relativement à l'utilisation de micro-ordinateurs. On présente également un résumé de l'utilisation du STIM par un nombre croissant de pays. De plus, le document décrit les dernières améliorations apportées au STIM de même que les projets futurs s'y rattachant. En conclusion, les auteurs expriment leurs opinions sur les tendances d'utilisation des micro-ordinateurs dans les domaines de la collecte et du traitement de données, et de l'utilisation et de la diffusion des données de recensement.

Les micro-ordinateurs comme outils de traitement des données de recensement

Grâce aux progrès technologiques des deux dernières années, le traitement d'importants volumes de données sur micro-ordinateurs est devenu encore plus pratique qu'auparavant. Les problèmes normalement associés aux tâches de traitement à grande échelle telles que les recensements de population disparaissent rapidement. Les unités de stockage des micro-ordinateurs sont plus nombreuses, plus rapides, plus fiables, plus petites, plus puissantes et moins coûteuses. Les vitesses de traitement des micro-ordinateurs ne cessent d'augmenter et la capacité de mémoire, de s'accroître.

Les bureaux de statistiques des pays en voie de développement ont découvert que cette technologie leur est plus accessible. Dans la plupart des cas, trouver un fournisseur local de micro-ordinateurs ne pose plus de problème. Habituellement, trois fournisseurs ou plus sont représentés. Ce marché plus compétitif a non seulement contribué à accroître le nombre d'options mais il a également réduit les prix et amélioré les conditions des contrats d'entretien.

Les micro-ordinateurs font désormais partie intégrante des opérations de la plupart des bureaux de statistiques des pays en voie de développement. On les utilise pour le traitement des enquêtes, la planification, l'établissement des budgets, le traitement de texte et bien d'autres tâches. Il y a des micro-informaticiens dans la plupart de ces pays et on retrouve sur le marché une variété de progiciels de micro-ordinateurs en différentes langues.

Les bureaux de statistiques ne se demandent plus s'ils doivent ou non utiliser les micro-ordinateurs; ils veulent maintenant savoir comment les utiliser de la meilleure façon. Ils se demandent quel type de logiciel et de matériel répondra le mieux à leurs besoins à court terme et à long terme. Ils cherchent à obtenir les ressources qui leur permettront d'acheter ce matériel et ces logiciels, de former leur personnel, d'assurer l'entretien du matériel et d'intégrer les micro-ordinateurs au contexte d'utilisation existant des ordinateurs principaux.

Les bureaux de statistiques connaissant mieux la micro-informatique, bon nombre d'entre eux choisissent d'utiliser les micro-ordinateurs pour traiter la totalité ou une partie de leurs données de recensement. Le Burkina Faso, le Sénégal et la Micronésie comptent parmi les pionniers ayant opté pour le traitement de leurs données de recensement à l'aide de micro-ordinateurs. Puis, le Niger, le Swaziland, le Yémen, le Honduras, l'Uruguay et la République Centrafricaine ont emboîté le pas. Les Philippines, le Pakistan et la Côte-d'Ivoire ont choisi de limiter l'utilisation des micro-ordinateurs à l'entrée des données provenant des questionnaires de recensement. D'autres pays, tels que la Tanzanie et le Cameroun ont décidé de répartir le traitement entre les micro-ordinateurs et un ordinateur principal.

Les micro-ordinateurs ont solutionné certains des problèmes éprouvés dans le passé par les pays qui n'utilisaient que des ordinateurs principaux pour traiter leurs données de recensement. Les pays ont pu s'offrir les capacités de traitement des micro-ordinateurs tout en respectant leurs budgets limités. La disponibilité de plusieurs micro-ordinateurs a résolu la plupart des problèmes d'accessibilité. L'entretien des micro-ordinateurs est plus facile et les normes relatives aux conditions ambiantes, moins limitatives. De plus, les exigences de formation sont moins strictes étant donné la disponibilité de logiciels faciles à apprendre et à utiliser même pour des personnes possédant peu d'expérience dans le traitement des données.

La disponibilité des micrologiciels qui comblent les besoins particuliers du traitement des données de recensement a grandement contribué à l'impact positif qu'ont eu les micro-ordinateurs dans les bureaux de statistiques. La plupart des pays mentionnés ci-dessus ont utilisé ou prévoient utiliser le STIM.

Système de traitement intégré sur micro-ordinateur : Origine et nouvelles caractéristiques

ORIGINE

L'introduction de micro-ordinateurs plus puissants au début des années 1980 ne fut pas accompagnée par des logiciels qui répondaient adéquatement aux exigences du traitement des recensements. L'ISPC a donc entrepris l'élaboration du STIM afin de fournir un logiciel pour micro-ordinateur pouvant exécuter les tâches principales liées au traitement des données de recensement : entrée de données, vérification des données, totalisation, analyse et contrôle opérationnel. Étant donné qu'un projet de cette ampleur nécessite énormément de temps et de ressources, l'ISPC a décidé d'utiliser le logiciel existant lorsque cela était possible et de fournir des produits intermédiaires pendant l'élaboration du STIM.

Les objectifs relatifs à la conception du STIM étaient les suivants : facilité d'utilisation, disponibilité d'un dictionnaire commun de données, capacité d'exécution avec une configuration standard de micro-ordinateurs et modularité. Le concept de modularité était particulièrement important au tout début de l'élaboration du STIM. Le système était conçu de manière à permettre la permutation facile des modules à mesure que des améliorations étaient apportées au logiciel. Par exemple, avant l'élaboration de l'élément d'entrée des données du STIM, CENTRY (Census Data ENTRY), on utilisait couramment des progiciels d'entrée de données disponibles sur le marché. Le remplacement par CENTRY ne nécessite pas d'effectuer des modifications aux autres modules d'application tels que vérification des données et totalisation. La souplesse d'emploi constitue un autre avantage de la modularité. L'utilisateur peut choisir les tâches qu'il désire exécuter à l'aide du STIM et celles qu'il veut effectuer avec d'autres logiciels et(ou) machines.

La première étape de l'élaboration du STIM consistait à rendre les progiciels CONCOR et CENTS 4 disponibles sur micro-ordinateurs. CONCOR (CONSISTENCY and CORrection) et CENTS 4 (CENsus Tabulation System, version 4) sont respectivement des programmes de

vérification des données et de totalisation élaborés par l'ISPC. Ces progiciels passent sur la plupart des ordinateurs principaux et en 1985, étaient utilisés par plus de 100 bureaux de statistiques dans les pays en voie de développement. Les deux progiciels étaient rédigés en langage COBOL afin d'être plus faciles à adapter aux différents ordinateurs principaux que l'on retrouvait dans les pays en voie de développement. La disponibilité de Realia COBOL, un compilateur COBOL puissant et efficace pour micro-ordinateurs, a facilité l'adaptation de CONCOR et CENTS 4 au cadre d'utilisation des micro-ordinateurs.

Les premières versions de CONCOR et CENTS 4 pour micro-ordinateurs sont apparues en 1984. Les jeux d'instructions pour ces versions étaient les mêmes que pour les versions destinées aux ordinateurs principaux. Cela signifiait que les mêmes programmes d'application CONCOR ou CENTS pouvaient être exécutés sur les ordinateurs principaux ou sur les micro-ordinateurs dans les bureaux de statistiques où l'on disposait de ces deux types de matériel. Un certain nombre de bureaux de statistiques, y compris ceux du Sénégal, du Cameroun et de la Somalie ont pu profiter de cette souplesse d'emploi.

L'obstacle suivant à franchir était la saisie des données, qui constitue souvent une goulet d'étranglement dans le traitement des données de recensement. L'ISPC a évalué des progiciels d'entrée de données produits sur le marché en vue d'identifier le meilleur logiciel pour la saisie des données d'enquête et de recensement. On a identifié deux progiciels qui répondaient aux critères de performance : ENTRYPOINT de Datalex Inc., et RODE/PC de DPX Inc. Au moment de l'évaluation, RODE/PC possédait un léger avantage par rapport à ENTRYPOINT aux niveaux du coût et de la vitesse. Par conséquent, l'ISPC a recommandé RODE/PC pour l'entrée de données de recensement sur micro-ordinateurs.

Westinghouse Public Applied Systems, sous le parrainage de l'Agence américaine de développement international (USAID), a adapté aux micro-ordinateurs les Computer Programs for Demographic Analysis (CPDA) (Programmes d'informatique pour analyse démographique), rédigés par le Bureau of the Census au début des années 1970, et très utilisés. Les Microcomputer Programs for Demographic Analysis (MCPDA) (Programmes de micro-informatique pour l'analyse démographique) constituent l'un des nombreux progiciels d'analyse démographique pour micro-ordinateurs. Actuellement, le Bureau of the Census travaille à l'élaboration d'un progiciel plus complet appelé Model Spreadsheets for Demographic Analysis (MSDA) (Modèle de feuilles de travail électronique pour l'analyse démographique). Il prépare également un manuel sur l'analyse des données de recensement. Ce manuel réunit les méthodes les plus utiles d'analyse démographique et de projection, y compris de brèves descriptions et discussions des avantages et des inconvénients des différentes méthodes ainsi que d'autres conseils précieux pour l'interprétation des résultats. Le manuel étudie également les logiciels disponibles pour exécuter chaque type d'analyse. Le MSDA sera disponible d'ici un an à peu près et le manuel, au cours de l'année suivante.

Le progiciel d'analyse statistique PC-CARP, élaboré par l'Université de l'état de l'Iowa, constitue l'élément d'analyse statistique du STIM. Au cours des deux dernières années, les bureaux de statistiques des pays en voie de développement ont grandement utilisé ce progiciel pour le calcul des variances dans les enquêtes, y compris les enquêtes postérieures au dénombrement. La Zambie, le Pérou, l'Égypte, le Zimbabwe, Costa Rica, Haïti et les Philippines figurent parmi les utilisateurs du PC-CARP. Grâce à son interface d'utilisateur, le PC-CARP est facile à utiliser pour les analystes ayant très peu ou aucune expérience du traitement des données.

Grâce à la disponibilité du micrologiciel pour l'entrée des données, la vérification des données, la totalisation, l'analyse démographique et l'analyse statistique, et à l'avènement des mémoires de grande capacité telles que les unités Bernoulli[2]/, il a été possible de recommander le traitement des données de recensement sur micro-ordinateurs aux pays

ayant une population de 10 millions ou moins. Le recensement de 1985 du Burkina Faso a été le premier recensement important entièrement traité sur micro-ordinateurs. On a utilisé RODE/PC, CONCOR, CENTS 4 et MCPDA. L'expérience du Burkina Faso est décrite de façon détaillée à la section sur les études de cas.

Depuis 1985, l'ISPC a ajouté de nouvelles options au STIM. Elles comprennent notamment un dictionnaire commun de données, un module d'entrée de données plus économique, des améliorations au CONCOR et au CENTS, un système de compte rendu et de gestion de recensement ainsi qu'un logiciel de planification des ressources de recensement.

NOUVELLES OPTIONS

Dictionnaire Commun de Données

À l'origine, CONCOR et CENTS 4 ont été élaborés comme progiciels autonomes pour les ordinateurs principaux. Pendant l'élaboration, on n'a prêté aucune attention à l'intégration des deux progiciels, les principales préoccupations étant la transférabilité et l'utilisation du logiciel sur des ordinateurs avec capacités de mémoire très limitées. L'utilisateur était donc obligé de définir le fichier de données pour le programme d'application CONCOR, puis pour le programme d'application CENTS 4, en utilisant chaque fois une notation différente. Une fois que CONCOR et CENTS 4 ont été adaptés aux micro-ordinateurs, la définition des données est devenue encore plus compliquée, car il fallait une troisième définition des données pour les progiciels commerciaux d'entrée des données tels que RODE/PC. Non seulement les multiples définitions de données prenaient du temps, mais elles étaient sujettes aux erreurs.

L'intégration de CONCOR et CENTS par l'entremise d'un dictionnaire commun de données a donné lieu à la version 3 de CONCOR et à la version 5 de CENTS, publiées au début de 1988. Avec ces versions, on ne définit qu'une seule fois le fichier de données devant être mis en forme et tabulé. Le module interactif du STIM, appelé DATADICT, demande à l'utilisateur d'entrer le nom et les caractéristiques de chaque catégorie de données. La définition qui en résulte, soit le dictionnaire de données, sert pour CONCOR et pour CENTS.

Dans ces nouvelles versions de CONCOR et de CENTS, l'interface de l'utilisateur a été considérablement simplifiée et les rapports d'édition de CONCOR sont plus concis. De nouvelles options ont été ajoutées à CONCOR telles que la capacité de consulter les catégories de données déjà définies et une plus grande souplesse d'emploi au niveau de la création de fichiers d'extraction.

CENTRY: logiciel pour l'entrée de données

Comme le coût des progiciels d'entrée de données retrouvés sur le marché augmentait, il devenait évident que les pays en voie de développement devaient trouver une solution moins coûteuse. Par exemple, l'entrée des questionnaires de pays ayant une population de 10 millions d'habitants nécessiterait pour un an approximativement 30 micro-ordinateurs, c'est-à-dire 30 copies de progiciel d'entrée de données. À environ $600 (US) la copie, le progiciel d'entrée de données devenait extrêmement cher. De plus, RODE/PC et ENTRYPOINT comportent des options telles que les contrôles logiques étendus qui sont essentielles à l'entrée des données d'enquête mais superflues pour l'entrée des données de recensement. Pour l'entrée des données de recensement, la vitesse est d'une importance capitale étant donné le volume considérable de données. Il est essentiel de saisir ce qui figure sur le questionnaire, que cela soit cohérent ou non. Les incohérences peuvent être corrigées plus tard par l'entremise de CONCOR mais ne doivent pas l'être par le personnel d'entrée de données. La plupart des pays en voie de développement ont des problèmes à atteindre des

taux d'entrée acceptables avec des systèmes dotés de contrôles logiques très limités ou sans contrôle logique.

Tout en tenant compte des besoins particuliers reliés à l'entrée des données de recensement, l'ISPC a élaboré un progiciel du nom de CENTRY comme module d'entrée de données du STIM. CENTRY est un progiciel dirigé par menus sur écran qui permet de réaliser des applications d'entrée de données, d'entrer et de vérifier les données et de recueillir des statistiques sur les opérations d'entrée de données. CENTRY utilise le même dictionnaire de données que CONCOR et CENTS. Les écrans de saisie programmables, le contrôle des valeurs correctes, la reproduction automatique des champs, le contrôle du curseur, le contrôle de la configuration de branchement, l'extraction et la modification d'enregistrements, et les statistiques relatives à l'opérateur en font un logiciel attrayant pour l'entrée des données de recensement. Les données entrées par CENTRY sont stockées dans un fichier ASCII prêt à être utilisé par d'autres progiciels tels que CONCOR. La facilité d'utilisation de CENTRY et du dictionnaire de données de STIM permet à toute personne possédant une formation minimale et connaissant le questionnaire de recensement de constituer une application d'entrée de données en une journée. CENTRY exige une mémoire de 256 kilo-octets et deux unités de disquettes. Il est rédigé en langage C afin de favoriser l'exécution rapide.

CENTRY sera publié en novembre 1988. Plusieurs pays dont la République Centrafricaine, le Burundi et les Philippines prévoient déjà l'utiliser pour l'entrée des données de recensement.

dCONTROL: système de gestion de recensement

Avec une opération de traitement de données aussi importante qu'un recensement national, il est necessaire de suivre chaque unité de données tout au long des diverses étapes de traitement. dCONTROL est un système interactif de contrôle et de gestion de recensement qui aide les gestionnaires de recensement à suivre le déroulement de ces étapes. Il sert également de fonction "d'enregistrement" pour les données classées par unité géographique telles que les secteurs de dénombrement (SD), empêchant ainsi la duplication ou l'omission de SD. Le dCONTROL permet également de produire des rapports de comptages préliminaires ainsi que d'autres rapports de gestion.

Dans son état actuel dCONTROL est un prototype pouvant être adapté aux besoins particuliers d'un pays. Il est utilisé au Niger et au Sénégal où le personnel de traitement des données des bureaux de statistiques l'a adapté à ses besoins spécifiques. L'ISPC travaille actuellement à l'élaboration d'une version générale du dCONTROL qui sera disponible en anglais, en français et en espagnol, au début de 1989.

dCONTROL exige un bon système cartographique définissant avec précision toutes les limites géographiques et administratives pour le recensement. Le système cartographique sert de base pour un code géographique qui identifie chaque territoire statistique jusqu'au niveau des SD. dCONTROL suit le déroulement des diverses étapes de traitement au niveau des SD. La première phase est habituellement la réception des questionnaires provenant des territoires par le bureau central; la dernière est l'entrée des données de recensement au clavier.

dCONTROL est rédigé en langage de programmation dBASE III PLUS compilé à l'aide du CLIPPER de Nantucket. Il passe sur un IBM PC/XT ou PC/AT ou sur un compatible doté d'une mémoire de 512 kilo-octets et d'une mémoire auxiliaire d'au moins 10 mégaoctets, habituellement sur disque dur. La quantité de mémoire requise sur disque dur dépend du nombre de SD dans le pays. Pour un pays comprenant 8,000 SD, des disques durs totalisant 2 mega-octets suffisent à mémoriser les programmes dCONTROL et les données.

CENPLAN

Bien que l'on ait réalisé beaucoup de progrès en très peu de temps au chapitre de l'élaboration de progiciels de traitement des données de recensement pour les micro-ordinateurs, il reste encore des améliorations à apporter. L'objectif consiste à rendre les résultats du recensement disponibles le plus tôt possible après le recensement. Un certain nombre de facteurs contribuent au retard des résultats de recensement, notamment la mauvaise planification.

L'ISPC élabore actuellement un progiciel appelé CENPLAN qui permet au responsable de la planification de recensement de déterminer les ressources requises pour le traitement du recensement en fonction de certains paramètres tels que la taille de la population, le matériel informatique disponible et les contraintes de temps. Bien que CENPLAN n'assure pas que la planification adéquate sera faite à temps pour le recensement, il fournit les outils de planification. CENPLAN utilise des feuilles de travail électroniques et sera disponible en français et en espagnol, ainsi qu'en anglais. La première version, qui porte uniquement sur le traitement informatique du recensement, sera disponible en novembre 1988.

Études de cas

Le nombre de pays utilisant le STIM pour le traitement des données de recensement continue de croître. Le tableau 1 présente une liste partielle des pays qui ont utilisé, utilisent ou prévoient utiliser le STIM pour traiter leurs données de recensement. Les bureaux de statistiques des pays en voie de développement tirent parti de la modularité du système qu'offre le STIM.

Certains pays, y compris le Niger, le Sénégal, le Bénin, le Burkina Faso, la République Centrafricaine, la Micronésie, le Burundi, la République Arabe du Yémen, le Mali, le Malawi, le Swaziland et les possessions américaines situées dans les régions éloignées du Pacifique, ont utilisé ou prévoient utiliser les micro-ordinateurs pour tous les aspects du traitement de leurs données de recensement. Le Sénégal, la République Arabe du Yémen et le Mali avaient prévu d'utiliser les micro-ordinateurs pour l'entrée de données et un ordinateur principal pour le reste du traitement des données. Toutefois, lorsqu'ils se sont rendus compte qu'ils pouvaient obtenir des résultats de manière plus rapide et plus rentable, ils ont décidé de traiter les données sur micro-ordinateurs. Ils se sont aperçus, par exemple, que CONCOR et CENTS s'exécutent en général plus rapidement sur micro-ordinateurs que sur des ordinateurs principaux. Même si la durée réelle d'exécution était théoriquement la même ou légèrement plus courte sur les ordinateurs principaux, ils se sont aperçus qu'ils obtenaient des résultats plus rapides sur les micro-ordinateurs spécialisés car ils n'avaient pas à partager les ressources informatiques avec d'autres utilisateurs.

D'autres pays comme la Somalie et la Côte-d'Ivoire ont élaboré leurs programmes CONCOR et CENTS en utilisant des micro-ordinateurs mais traitent les données sur des ordinateurs principaux. D'autres pays comme le Ghana et l'Éthiopie entraient les données sur micro-ordinateurs, puis téléchargaient les données dans un ordinateur principal pour le traitement. Par ailleurs, le Honduras, les Philippines et la Tanzanie n'ont pas encore pris de décision à savoir s'ils utiliseront les micro-ordinateurs, un ordinateur principal ou une combinaison des deux pour traiter les données de recensement. Les critères de sélection ne sont pas nécessairement techniques. Certains pays doivent utiliser les ordinateurs principaux pour le traitement des données de recensement parce que leur gouvernement ou leurs donateurs ont acheté un ordinateur principal il y a plusieurs années aux fins express d'utilisation pour le recensement. Malgré la révolution technologique, il est nécessaire d'amortir cet équipement avant que les preneurs de décision puissent justifier l'achat de nouvel équipement.

La conception modulaire du STIM permet aux bureaux de statistiques des pays en voie de développement de profiter du matériel et des logiciels existants appropriés à l'entrée de données de recensement. Par exemple, la Tanzanie et le Cameroun utilisent des mini-ordinateurs tels que le ICL/DRS 300 et le System/36 d'IBM pour l'entrée de données. La République Arabe du Yémen utilise les micro-ordinateurs pour l'entrée de données mais a imparti l'élaboration de ses programmes d'entrée de données en raison de l'entrée de données spécialisées en caractères arabes. Tous ces pays prévoient utiliser ou utilisent actuellement CONCOR et CENTS pour la vérification et la totalisation des données.

En outre, comme d'autres éléments du STIM deviennent disponibles, les bureaux de statistiques des pays en voie de développement peuvent planifier de les exploiter au maximum. Par exemple, le Burundi, la République Centrafricaine et les Philippines prévoient utiliser CENTRY pour l'entrée de leurs données de recensement.

Bien que le STIM ait été principalement conçu pour le traitement des données de recensement, on l'utilise également pour le traitement des données d'enquête. La Gambie, le Maroc, la Zambie, le Sénégal, le Portugal, le Ghana, le Ruanda, l'Égypte et les Samoa Américaines utilisent un ou plusieurs modules du STIM pour leurs enquêtes.

Les deux études de cas suivantes, sur le Burkina Faso et le Sénégal, ont été exposées dans le document de Toro et Melaney afin d'illustrer l'effet positif du STIM sur les activités de traitement de données de recensement. Une mise à jour de leurs expériences suit. Les États Fédérés de la Micronésie (EFM), objet de la troisième étude de cas présentée dans le document de Toro et Melaney, ont continué d'utiliser les micro-ordinateurs avec succès pour le traitement de données de recensement dans certains de leurs États. Les EFM songent actuellement à élaborer une méthode de recensement national qui standardiserait les données, les questionnaires ainsi que les activités de traitement pour tous les États.

Burkina Faso

En décembre 1985, l'Institut National de la Statistique et de la Démographie (INSD) du Burkina Faso, anciennement la Haute-Volta, effectuait son recensement national. Il ne lui a fallu que 18 mois pour entrer et traiter les données sur la population, évaluée à 9 millions d'habitants.

Trois micro-ordinateurs IBM PC/AT dotés d'une mémoire de 512 kilo-octets et de disques durs d'une capacité de mémoire de 20 méga-octets ont été utilisés pour traiter les données. Le matériel d'entrée de données était composé de 22 PC d'IBM ayant chacun une mémoire de 512 kilo-octets et de deux IBM PC/XT avec disque dur ayant une capacité de mémoire de 10 méga-octets et une mémoire de 640 kilo-octets. Les périphériques s'y rattachant comprenaient six imprimantes, deux dérouleurs de bandes magnétiques DIGIDATA et trois unités Bernoulli avec cartouches amovibles de 10 méga-octets. Les dérouleurs de bandes magnétiques étaient utilisés aux fins de sauvegarde et d'archivage tandis que les unités Bernoulli servaient de mémoire centrale pour les données de recensement. On a utilisé les progiciels RODE/PC, CONCOR et CENTS 4 respectivement pour l'entrée de données, la vérification des données, et la totalisation.

Le personnel de l'INSD chargé du traitement de données pour le projet de recensement comptait deux programmeurs à temps plein et un conseiller à long terme envoyé par l'ONU. De plus, en vertu d'une entente avec le USAID, le Bureau of the Census fournissait de l'aide en matière de traitement de données sous forme de formation, d'élaboration de programmes et de surveillance du programme d'application d'entrée de données RODE/PC, du programme de vérification des données CONCOR et du programme de totalisation CENTS 4.

Le rapport des comptages préliminaires reposant sur les comptages manuels était disponible en mars 1986. Deux équipes de 25 opérateurs commençaient à entrer les données de l'échantillon de 15 pour cent en septembre 1986. Deux mois plus tard, cette opération de saisie des données était finie, et en mars 1987, les totalisations pour l'échantillon de 15 pour cent étaient produites. Les données d'échantillonnage préparées occupent 11 cartouches Bernoulli de 10 méga-octets chacune. (Il est à noter qu'on a commencé l'entrée de données avec plusieurs mois de retard, étant donné l'arrivée tardire des micro-ordinateurs).

L'entrée du reste des données a eu lieu de décembre 1986 à mai 1987. Les 40 tables pour l'ensemble du pays étaient disponibles en mars 1988, soit 18 mois après le début de l'entrée de données. Les données entrées pour chaque secteur de dénombrement étaient stockées sur une disquette. On n'utilisait jamais plus de 600 disquettes à la fois étant donné qu'elles étaient recyclées après la vérification des données. Une fois que les données avaient été complètement vérifiées, tous les SD d'une province (200 SD en moyenne par province) étaient transférés sur une ou plusieurs cartouches Bernoulli. Environ 10,000,000 d'enregistrements (9,000,000 pour la population et 1,000,000 pour le logement) de 52 caractères de longueur chacun furent sauvegardés. Soixante cartouches Bernoulli de 10 méga-octets chacune furent utilisées pour stocker une copie des données pour l'ensemble du pays. On a conservé cinq versions de ces données.

Ces versions comprennent les données brutes au niveau des secteurs de dénombrement (format de fichier RODE/PC et format ASCII), les données regroupées au niveau provincial et deux copies des données préparées.

Un système d'étiquettes externe simple a assuré la gestion efficace d'environ 600 disquettes et 240 cartouches Bernoulli. Chaque disquette était clairement identifiée par une étiquette externe indiquant le secteur de dénombrement. Chaque cartouche Bernoulli était identifiée par une étiquette indiquant une province. La gestion et le traitement des données de recensement constituaient des tâches réalisables mais plutôt difficiles étant donné que les cartouches Bernoulli ne pouvaient contenir que 10 méga-octets de données à la fois et que les micro-ordinateurs IBM PC/AT avaient une mémoire sur disque dur de seulement 20 méga-octets. En raison de problèmes techniques, les dérouleurs de bandes magnétiques ne pouvaient servir qu'à l'archivage et non au traitement des données de recensement.

Tableau 1. Liste partielle des utilisateurs du STIM pour les recensements de la population

Pays	Date du recensement	Pop. évaluée (millions)	Contrôle opérationnel	Entrée de données	Vérification	Tabulation
Bénin	1989	4.3	manuel	CENTRY	CONCOR	CENTS
Burkina Faso	12/85	9	part. aut.	RODE/PC	CONCOR	CENTS
Burundi	8/90	5	non décidé	CENTRY	CONCOR	CENTS
Cameroun	4/87	10.3	manuel	IBM S/36	CONCOR	CENTS
République centrafricaine	10/88	2.7	dCONTROL	CENTRY	CONCOR	CENTS
Comores	1990	0.5	manuel	non décidée	CONCOR	CENTS
Côte-d'Ivoire	3/88	10	manuel	RODE/PC	CONCOR/a	CENTS/a
Honduras	5/88	4.8	dCONTROL	KeyEntry III	CONCOR/b	CENTS/b
Malawi	9/87	7.4	manuel	ICL mini	CONCOR	CENTS
Mali	4/87	8.4	manuel	RODE/PC	CONCOR	CENTS
Mauritanie	10/87	1.9	manuel	non connue	CONCOR	CENTS
Micronésie (Pohnpei)	9/85	0.3	manuel	Entrypoint	CONCOR	CENTS
Niger	5/88	7.5	dCONTROL	RODE/PC	CONCOR	CENTS
Philippines	1990	57	non décidé	CENTRY	CONCOR/b	CENTS/b
Sénégal	5/88	7.5	dCONTROL	RODE/PC	CONCOR	CENTS
Somalie	11/86	7.8	manuel	prog. pers.	CONCOR/a	CENTS/a
Swaziland	8/86	0.7	manuel	RODE/PC	CONCOR	CENTS
Tanzanie	8/88	23.5	manuel	ICL mini	CONCOR/b	CENTS/b
République arabe du Yémen	2/86	9	part. aut.	prog. pers.	CONCOR	CENTS
É.-U. (régions isolées du Pacifique)	4/90	0.3	manuel	CENTRY	CONCOR	CENTS

a/ Élaboration de programmes faite sur micro-ordinateurs; traitement de la production sur des ordinateurs principaux.

b/ Élaboration de programmes faite sur micro-ordinateurs; n'ont pas encore décidé ce qui sera utilisé pour le traitement de la production.

Avec CONCOR et CENTS, le fichier complet sur le recensement de la population n'a pas à être traité ou trié en tant qu'unité. Les données peuvent être vérifiées lot par lot en utilisant CONCOR. La taille des lots peut varier. Les lots de données vérifiées (ou nettoyées) peuvent être fusionnés en lots plus importants (ou unités géographiques) pour la tabulation à l'aide de CENTS. Les tableaux croisés qui en résultent peuvent être sauvegardés dans des fichiers plus petits, puis regroupés de manière à produire des tables pour les entités géographiques plus importantes. Par conséquent, dans un système utilisant plus d'un micro-ordinateur, plusieurs machines peuvent traiter des régions géographiques différentes simultanément, ce qui réduit la durée globale de traitement.

Dans le cas du Burkina Faso, la vérification et la totalisation étaient effectuées au niveau de la province. Le programme CONCOR prenait 20 minutes en moyenne pour la vérification des données de chacune des 30 provinces. L'exécution des programmes de tabulation CENTS prenait environ 15 minutes par province. Les tabulations intermédiaires pour les provinces ont été fusionnées par l'entremise du CENTS afin de produire des tables pour le pays. On s'est servi de trois IBM PC/AT dotés de mémoire de 20 méga-octets pour la vérification des données et totalisation.

L'entrée de données pour tout le pays a pris 9 mois au lieu des 12 mois prévus, et ce, principalement en raison du rythme accéléré d'entrée atteint. La cadence de frappe variait de 9,000 à 14,000 touches à l'heure avec un taux d'erreur de .3 pour cent. Comme l'INSD n'était pas en mesure d'offrir de stimulants pécuniers au personnel chargé de l'entrée, il leur a fourni d'autres types de motivation. Par exemple, le personnel le plus efficace pouvait choisir de travailler pendant le quart (matin ou soir) qui lui convenait le mieux et planifier ses congés plus librement. Il leur était également promis un travail permanent à l'INSD après le recensement. Étant donné le taux élevé de chômage au Burkina Faso, la plupart des employés étaient motivés simplement par le fait de conserver un emploi relativement bien rémunéré. Sur les 500 candidats ayant passé un examen écrit et une entrevue, on n'en a choisis que 50. Avant tout, les personnes chargées de la saisie de données étaient travailleuses et fières de leur travail.

L'entretien des micro-ordinateurs était effectué par l'équipe de traitement de données. Un stock de pièces de rechange leur a permis de maintenir le temps d'arrêt à 4 pour cent. Ce sont les unités de disquettes et les claviers qui ont causé le plus de problèmes. Malgré cela et malgré certains problèmes électriques initiaux, l'entretien du matériel s'est fait avec à propos, sans perturber les activités.

En tant que pionnier, l'INSD a démontré que le traitement du recensement d'une population de 9,000,000 d'habitants à l'aide de micro-ordinateurs est non seulement réalisable et rentable, mais qu'il constitue également une méthode efficace pour produire des résultats dans des délais raisonnables.

Sénégal

La Direction de la Statistique (DS) a effectué le deuxième recensement de la population nationale de la République du Sénégal en mai 1988. La population était évaluée à 7,500,000 d'habitants.

Devant le coût raisonnable du matériel de micro-informatique, la disponibilité des logiciels adéquats d'entrée de données et les possibilités d'utilisation des micro-ordinateurs pour d'autres projets une fois la saisie de données terminée, la DS a décidé d'acheter des micro-ordinateurs pour l'entrée de données au lieu de louer l'équipement IBM 3742. La vérification et la totalisation des données devaient être effectuées sur un IBM 370/145, au Centre informatique du ministère des Finances, un centre de production de données utilisé par d'autres agences gouvernementales.

La configuration micro-informatique originale comprenait 18 PC d'IBM dotés de 256 kilo-octets de mémoire pour l'entrée de données, deux IBM PC/XT avec 10 méga-octets de mémoire sur disque dur pour l'élaboration de programmes, cinq imprimantes et un dérouleur de bandes magnétiques utilisé aux fins de sauvegarde des données. On a utilisé des unités d'alimentation non interruptibles afin de protéger l'alimentation électrique.

On a choisi le progiciel RODE/PC pour l'entrée de données. Les données devaient être vérifiées à l'aide de CONCOR et totalisées par CENTS 4. Les programmes CONCOR et CENTS 4 pouvaient être élaborés à l'aide des micro-ordinateurs puis transférés sur l'ordinateur principal pour les phases d'exécution.

Lors d'une visite d'assistance technique de trois semaines à Dakar en janvier 1986, dans le cadre d'une entente avec USAID, des conseillers en matière de traitement de données de l'ISPC ont dirigé un atelier de CENTS 4, installé CONCOR et CENTS 4 sur le processeur central IBM et discuté des plans de totalisation et des spécifications avec le personnel de la DS.

À l'origine, le personnel de traitement de données responsable du déroulement du recensement était composé du chef de la section de traitement de données et du chef de la section d'entrée de données de la DS. Ceux-ci ont poursuivi leur formation pendant une visite d'étude - travail à l'ISPC, à Washington, D.C., au début de 1986. Au cours de ce séjour de six semaines, le superviseur de l'entrée de données, qui n'avait aucune expérience en programmation, a appris à utiliser le RODE/PC dirigé par menus et a participé au développement de l'application d'entrée de données pour le questionnaire de recensement. Le chef du traitement de données, une analyste de système expérimentée ayant déjà reçu une formation sur CONCOR et CENTS 4, est demeurée à l'ISPC pendant trois mois. En plus d'apprendre le fonctionnement du RODE/PC, elle a rédigé les programmes de vérification de données et d'imputation de CONCOR. Elle a également effectué la majeure partie de la programmation de totalisation à l'aide de CENTS 4. Tout le travail a été effectué sur un IBM PC/XT.

À leur retour au Sénégal, ils étaient facilement en mesure d'apporter des modifications aux programmes lorsque des changements importants étaient apportés à certaines parties du questionnaire. Les données du recensement pilote, effectué en mars 1987, ont été entrées à l'aide de micro-ordinateurs. Les programmes CONCOR et CENTS 4 ont été transférés sur l'ordinateur principal. Il n'a pas été nécessaire de changer le code. Les données du recensement pilote introduites ont également été transférées des disquettes de 5¼ pouces aux disquettes de 8 pouces afin d'en faciliter le transfert sur l'ordinateur principal.

Quoiqu'il n'ait rencontré aucun problème majeur pendant le traitement de données du recensement pilote sur l'ordinateur principal, le personnel du traitement de données fut frustré par certains inconvénients. En effet, il devait transporter les données à l'emplacement de l'ordinateur principal situé à plusieurs milles du lieu de travail. Après avoir effectué un test, on s'est rendu compte qu'en raison de la nécessité de partager les installations informatiques, il fallait plus de temps pour passer les programmes CONCOR et CENTS 4 sur le processeur central que sur les micro-ordinateurs spécialisés. Le personnel de traitement de données a également trouvé que l'interaction était plus facile avec le PC d'IBM qu'avec l'ordinateur principal d'IBM.

Devant les nombreuses expériences positives avec les micro-ordinateurs, le personnel du traitement de données a décidé de traiter les données de l'ensemble du recensement sur des micro-ordinateurs plutôt que sur des ordinateurs principaux. Afin d'augmenter la configuration actuelle des micro-ordinateurs de manière à permettre le traitement sur place des données du recensement, le personnel de traitement de données a fait l'acquisition de

mémoires de grande capacité qui consistaient en trois cartouches Bernoulli de 20 méga-octets et un dérouleur de bandes magnétiques. Deux de ces cartouches Bernoulli contiennent un disque dur doté d'une mémoire de 80 méga-octets. Bien que la gestion de fichiers sur disquettes, cartouches Bernoulli et bandes magnétiques constitue un élément crucial du traitement, le personnel a jugé que les problèmes de gestion de fichiers étaient moins importants que les inconvénients reliés au travail avec un ordinateur principal. De plus, le chef de la section d'entrée de données ne s'inquiétait pas au sujet du traitement des fichiers de données sur le recensement étant donné qu'il avait élaboré et mis en application le système de gestion de supports magnétiques utilisé lors du recensement précédent de la population.

La DS a acquis six micro-ordinateurs IBM PC/AT supplémentaires. Quatre sont utilisés par le personnel du traitement de données pour l'élaboration de programmes et pour le traitement de la production. Les deux autres micro-ordinateurs seront utilisés par les démographes et les statisticiens pour l'analyse des données de recensement. La DS prévoit utiliser le MCPDA pour l'analyse démographique des données de recensement.

Le chef de la section de traitement de données a transmis son expertise en micro-informatique à trois analystes de système qui travaillent également au traitement de données de recensement. De plus, l'ISPC fournit à la DS une assistance technique pour tous les aspects des activités de traitement de données de recensement. Un de ces aspects est la surveillance des activités de traitement de données de recensement et la production de rapports de gestion y compris le rapport de comptages préliminaires. Le personnel de la DS a adapté le prototype dCONTROL aux besoins spécifiques du Sénégal. Les rapports de comptages préliminaires qui reposent sur les comptages manuels du dénombrement de mai et juin 1988 étaient disponibles dès octobre 1988. La base de données et les programmes dCONTROL occupaient moins de 2 méga-octets de mémoire.

La DS prévoit traiter en premier lieu un échantillon de 10 pour cent des données. Toutefois, la disponibilité d'un emplacement permanent pour ranger les questionnaires, coder et entrer les données a retardé le début du traitement de données de recensement jusqu'à la mi-septembre. Le programme d'entrée de données RODE/PC et les programmes de vérification de données CONCOR ont été testés. On en est à la version finale des programmes de totalisation CENTS. On a élaboré un système de contrôle du rendement de l'opérateur d'après les statistiques sur l'opérateur telles que la cadence de frappe à l'heure et le taux d'erreurs de frappe. Ce système dBASE III PLUS, produit des rapports avec des statistiques relatives aux opérateurs d'après l'information contenue dans le fichier ASCII produit par le programme d'entrée de données RODE/PC.

Entre-temps, le personnel de la DS utilise RODE/PC et CENTS pour d'autres enquêtes. L'enthousiasme du personnel de la DS à l'égard de l'utilisation des micro-ordinateurs va même plus loin. Dans le cadre de l'étape d'analyse et de publication, la DS prévoit utiliser les micro-ordinateurs pour la cartographie thématique. Il s'agit de la création informatisée de cartes reflétant les disparités régionales en ce qui a trait aux caractéristiques de la population. Elle requiert des logiciels de cartographie ainsi que du matériel périphérique tel qu'un convertisseur analogique/numérique et un traçeur. Le U.S. Bureau of the Census assurera la formation.

Au cours des trois dernières années, la DS s'est montrée plus confiante et plus enthousiaste à l'égard des micro-ordinateurs. Non seulement utilise-t-on les micro-ordinateurs pour traiter les données de recensement mais on les utilise également pour plusieurs autres projets et enquêtes. La plupart des analystes de système possèdent un micro-ordinateur sur leur bureau et donnent une formation en micro-informatique aux programmeurs juniors. Les démographes et les statisticiens utilisent également les micro-ordinateurs pour leur travail.

La compétence sur le sujet acquise lors du dernier recensement, les connaissances accrues en micro-informatique et l'enthousiasme du personnel, laissent entrevoir un avenir favorable pour le traitement des données de recensement.

Pronostic

Les expériences du Burkina Faso et du Sénégal illustrent bien les répercussions positives des micro-ordinateurs et du STIM sur les activités de traitement de données de recensement de nombreux pays. Il y a trois ans, ces pays faisaient office de pionniers dans l'utilisation des micro-ordinateurs pour les recensements. Maintenant, plus de 18 pays utilisent ou prévoient utiliser les micro-ordinateurs pour un aspect quelconque du traitement de leurs données de recensement. Ce nombre croît rapidement en même temps que les bureaux de statistiques et les donateurs prennent conscience des avantages des micro-ordinateurs.

Le contrôle opérationnel et la gestion de fichiers représentent toujours un défi. À mesure que les densités d'enregistrement des mémoires de grande capacité continuent d'augmenter, le contrôle opérationnel et la gestion de fichiers deviennent plus faciles, peut-être même davantage que sur les ordinateurs principaux. Le système de compte rendu et de gestion de recensement dCONTROL aide à relever ce défi.

Malgré toutes ces améliorations, la saisie de données constitue toujours l'opération la plus longue du traitement de données de recensement. La vitesse à laquelle les données peuvent être introduites est limitée, peu importe la rapidité du logiciel. Bien que l'on puisse ajouter des postes de travail afin de diminuer le temps total d'entrée, une telle opération entraîne des problèmes de logistique. Comme la technologie relative aux lecteurs optiques de marques ne cesse de progresser, d'ici cinq ans, elle devrait offrir aux pays en voie de développement une solution de rechange rentable en ce qui a trait à l'entrée de données.

La disponibilité d'outils plus rentables pour le traitement de données de recensement devrait permettre aux bureaux de statistiques des pays en voie de développement de concentrer leur attention et leurs ressources sur l'utilisation et la diffusion des données de recensement. Les progiciels pour la production de cartes thématiques et l'édition électronique, entre autres, permettront d'atteindre cet objectif. De plus, les améliorations apportées au disque optique facilitent l'accès et le transfert d'importants volumes de données.

Les utilisateurs des données de recensement eux-mêmes disposent de plus d'outils, ce qui devrait, en principe, alléger les responsabilités des bureaux de statistiques en tant que principaux diffuseurs de données de recensement. Comme bon nombre d'utilisateurs de données de recensement n'ont jamais pu obtenir les résultats de recensement ou les ont obtenus cinq ans ou plus après le dénombrement, ce qui rendait leur interêt pratiquement nul, la disponibilité de toute donnée représente une amélioration importante. Les utilisateurs seront heureux d'avoir accès à un sous-ensemble de données de recensement sur disquettes, qu'ils pourront utiliser sur leurs propres micro-ordinateurs. Espérons que les bureaux de statistiques et les utilisateurs se concentreront sur l'analyse et la diffusion de données sûres et ne se laisseront pas distraire par les nombreux choix de présentation qu'offre la nouvelle technologie.

Les besoins en traitement de données et les logiciels changent avec l'évolution de la technologie informatique. Le STIM est le fruit de nombreuses années d'expérience du personnel de l'ISPC et la réponse aux demandes des gestionnaires de traitement de données de recensement des pays en voie de développement. Si les ressources le permettent, l'ISPC continuera à appuyer et à modifier les divers modules du STIM en observant attentivement son utilisation par les bureaux de statistiques des pays en voie de développement, en prêtant l'oreille aux suggestions d'améliorations des utilisateurs du STIM et en profitant pleinement

du contexte d'utilisation de micro-ordinateurs disponibles dans les bureaux de statistiques des pays en voie de développement.

L'objectif principal de l'ISPC est de rendre ces progiciels utilisables par des personnes qui ne sont pas des programmeurs qualifiés ou expérimentés. Au cours des prochaines années, l'ISPC apportera des modifications au CONCOR et au CENTS en ce qui a trait à la facilité d'utilisation. Un des buts est de rendre interactifs les aspects définitionnels de ces progiciels de façon que les utilisateurs n'aient pas à apprendre un langage procédural pour les utiliser. Certains aspects des progiciels tels que la définition des contrôles de cohérence complexes sont mieux énoncés par l'entremise d'un langage procédural. D'autres, comme la définition des formats des tables, se font mieux de façon interactive. L'ISPC élaborera également des guides d'initiation et des exemples détaillés de recensement afin de faciliter le processus d'apprentissage.

L'enthousiasme pour la nouvelle technologie ne devrait pas amoindrir l'effet de la planification, des politiques et de la communication sur le traitement d'un recensement. Après tout, les micro-ordinateurs ne sont que des outils mis à la disposition des gens; ils ne remplacent pas une planification soignée ni une coordination étroite.

Conclusions

Avec les expériences des pays en voie de développement tels que le Burkina Faso et le Sénégal, l'utilisation des micro-ordinateurs pour le traitement de recensements de dix millions de personnes est devenue une réalité. Grâce aux progrès technologiques enregistrés au cours des trois dernières années, certains des problèmes normalement associés au traitement à grande échelle disparaissent rapidement. La disponibilité d'outils rentables pour le traitement de données de recensement devrait permettre aux bureaux de statistiques des pays en voie de développement de concentrer leur attention et leurs ressources sur l'utilisation et la diffusion de données de recensement. Nous attendons avec impatience les répercussions des progrès technologiques futurs sur le traitement des données à grande échelle.

REMARQUES

1/ Dans le présent document, toute référence aux ordinateurs principaux comprend également les mini-ordinateurs.

2/ Les unités Bernoulli sont des systèmes de disque dur avec cartouches amovibles. Elles peuvent servir à la fois de mémoire à grande capacité et de système de sauvegarde. Leur densité de stockage de l'information est plus grande et elles permettent une plus grande rapidité d'accès aux données. Chaque cartouche est dotée d'un ou de deux lecteurs de cartouches. Selon le modèle, les unités peuvent lire des données à partir de cartouches de 10 ou de 20 méga-octets. Comme ces cartouches sont amovibles, elles offrent une capacité de mémoire pratiquement illimitée. Les nouveaux modèles sont également dotés de disques durs avec une mémoire de 80 méga-octets en plus des cartouches amovibles. En outre, ils résistent très bien aux mauvais traitements qui entraîneraient la destruction ou la perte de données sur des disques durs. Les unités Bernoulli sont utilisées depuis plusieurs années dans les pays en voie de développement et se sont avérées très fiables.

REFERENCES

Anderson, Patricia, and Ondra, Thomas. 1987. Trip and assessment report to Pohnpei, Federated States of Micronesia. U.S. Bureau of the Census.

Chamberlain, Kathleen. 1988. Trip reports to Mogadishu, Somalia. U.S. Bureau of the Census.

Chamberlain, Kathleen. 1988. Trip reports to Dar Es Salaam, Tanzania. U.S. Bureau of the Census.

Gomez, Enrique. 1988. Trip reports to Tegucigalpa, Honduras. U.S. Bureau of the Census.

Gomez, Enrique, et al. 1987. Trip reports to Tegucigalpa, Honduras. U.S. Bureau of the Census.

Hie, Jean-Marc. 1988. Trip report to Ouagadougou, Burkina Faso. Economic Commission for Africa. United Nations.

Le, Thanh, and Toro, Vivian. 1988. Trip report to Dakar, Senegal. U.S. Bureau of the Census.

Le, Thanh. 1988. Trip reports to Bujumbura, Burundi. U.S. Bureau of the Census.

Megill, David, and Toro, Vivian. 1988. Trip report to Niamey, Niger. U.S. Bureau of the Census.

Megill, David. 1988. Trip report to Niamey, Niger. U.S. Bureau of the Census.

Melaney, Thomas, and Toro, Vivian. 1988. Trip report to Dakar, Senegal. U.S. Bureau of the Census.

Rowland, Sandra, and Melaney, Thomas. 1988. Trip report to Bujumbura, Burundi. U.S. Bureau of the Census.

Sawaya, Selma, and Banerjee, Kakali. 1987. Trip report to Abidjan, Cote d'Ivoire. U.S. Bureau of the Census.

Stroot, Michael. 1988. Trip report to Sana'a, Yemen Arab Republic. U.S. Bureau of the Census.

Toro, Vivian, et al. 1987. Trip report to Niamey, Niger. U.S. Bureau of the Census.

Toro, Vivian, et al. 1987. Trip report to Dakar, Senegal. U.S. Bureau of the Census.

Toro, Vivian, and Melaney, Thomas. 1987. "The Use of Microcomputers for Census Processing in Developing Countries," paper presented at the American Statistical Association Meetings, San Francisco, California.

U.S. Bureau of the Census. 1988. "Population Census Support Status of Sub-Saharan Africa"

UN RAPPORT SUR LE TRAITEMENT D'ENQUÊTE RÉUSSI DANS DES PAYS EN VOIE DE DÉVELOPPEMENT: L'APPROCHE DES MICRO-ORDINATEURS AU NIVEAU DES ENQUÊTES DÉMOGRAPHIQUES ET DE SANTÉ

J. CUSHING,

Institute for Resource Development/Westinghouse

1. Introduction

Le programme d'enquêtes démographiques et relatives à la santé (PEDS), actuellement réalisé par le Institute for Resource Development-Westinghouse (IRD), est financé par la U.S. Agency for International Development (US-AID). La première phase quinquennale du PEDS, entamée en octobre 1984, a fourni l'aide technique et financière nécessaire à la réalisation de 35 enquêtes menées dans 30 pays en voie de développement. La seconde phase du programme, comprenant 25 enquêtes supplémentaires, fut récemment attribuée à l'IRD; le programme pourra ainsi se poursuivre pour encore cinq années consécutives.

Les enquêtes menées par le PEDS ont eu des échantillons allant de 3,000 à plus de 12,000 répondants. Les questionnaires de base, au nombre de deux, comprennent une liste de ménages et un questionnaire individuel de huit sections s'appliquant aux femmes éligibles figurant sur la liste. Chaque questionnaire recueille des données sur les caractéristiques fondamentales de la personne interviewée et de son conjoint (si opportun), la fécondité (incluant toutes les naissances), la connaissance et la pratique de la contraception, le nombre d'enfants désiré, la santé de la mère et des enfants nés au cours des cinq dernières années, ainsi que la taille et le poids des enfants âgés de 3 à 36 mois. Les deux questionnaires de base diffèrent en ce que le premier met l'accent sur la pratique de la contraception alors que le second met l'accent sur la santé. Une table à deux dimensions contenant les données sur une personne interviewée comporte environ 2,000 caractères.

La première enquête du PEDS fut menée au Salvador au début de 1985. Afin d'obtenir un fichier pour les tabulations préliminaires, un sous-ensemble de variables fut encodé sur des formules, saisi par lecteur optique sur micro-ordinateur, écrit sur l'unité centrale et tabulé. L'enquête fut entièrement traitée au Salvador avec des méthodes traditionnelles comprenant la transcription des données, l'entrée des données de façon non-interactive et la vérification cyclique des données par lots, tel qu'indiqué à la figure 1.

Le traitement par micro-ordinateur avait été envisagé pour le PEDS au Salvador, mais cette option fut rejetée parce que ses possibilités n'avaient pas encore été mises à l'épreuve dans le cadre d'enquêtes à grande échelle telles que celles réalisées par le PEDS. De plus, les micrologiciels pour le traitement des données étaient encore à l'étude. À l'automne 1985, le PEDS dut par contre obligatoirement utiliser le traitement des données sur micro-ordinateur, car son échéancier ne lui permettait plus les retards qui se présentent normalement au cours du traitement des données sur unité centrale dans les pays en voie de développement. On utilisa les logiciels qui existaient alors sur le marché, et le PEDS élabora un système intégré permettant d'éviter les goulots d'étranglement qui s'étaient produits lors d'enquêtes antérieures en pays en voie de développement.

Grâce à l'effort alors réalisé, le PEDS termine maintenant la saisie des données environ deux semaines après la fin du travail sur le terrain. À ce stade, le fichier de données est bien structuré, il ne contient aucune erreur d'intervalle ou de routine: "passer à" et il reste peu d'incohérences dans les questionnaires. La vérification secondaire par lots des cas de cohérence interne complexe peut se poursuivre pour encore un mois ou deux. On produit maintenant des rapports préliminaires en moins de trois mois après le travail sur le terrain. Les retards des rapports finaux ne sont plus attribuables au traitement des données mais plutôt au temps requis pour réviser, approuver et produire le rapport. On publie néanmoins les rapports finaux environ un an après que le travail sur le terrain soit terminé.

Le tableau 1 montre de façon cumulative le travail réalisé au cours de la Phase 1 du PEDS. Il est à noter qu'il n'y a pas eu de rapport préliminaire pour 4 enquêtes; que 2 enquêtes étaient de nature expérimentale, exigeant une préparation de rapport plus longue; et que 5 enquêtes sont encore sur le terrain. Les chiffres entre parenthèses sont des estimations pour la fin de 1988 et l'année 1989.

Tableau 1
Nombre cumulatif d'enquêtes du PEDS et de rapports par année

Activité	1985	1986	1987	9/1988	1988	1989
Enquêtes menées sur le terrain	1	10	23	29	(34)	(35)
Rapports préliminaires publiés	0	1	12	20	(22)	(31)
Premiers rapports de pays publiés	0	0	3	11	(21)	(35)

En examinant le tableau 1 on constate que la production des rapports pour les premières enquêtes (jusqu'à la fin 1986) fut considérablement plus lente que la production actuelle. Les rapports préliminaires pour quelques enquêtes commencées au début de 1986 ne furent accessibles qu'en 1987. Les rapports préliminaires d'enquêtes menées en 1988 furent cependant produits dans une période de trois mois suivant l'enquête réalisée sur le terrain. Les rapports finaux de toutes les enquêtes effectuées en 1987 ont été publiés ou seront publiés avant la fin de 1988.

En moyenne, le travail sur le terrain a duré trois ou quatre mois par enquête. Les rapports préliminaires furent publiés trois mois plus tard en Amérique latine, en Afrique du Nord et en Asie, et cinq mois plus tard en Afrique sous-saharienne. La plupart des tout premiers rapports sur des enquêtes nationales furent publiés environ un an après la fin du travail sur le terrain; certains ont pris de dix-huit mois à deux ans. On peut comparer ces données avec ce qui se produisait pour les enquêtes antérieures, dont les résultats n'étaient souvent publiés que deux à quatre ans plus tard.

La diminution du décalage entre le travail sur le terrain et la publication des rapports peut être attribuée à plusieurs facteurs: le passage de la grande informatique à la micro-informatique dans l'organisme d'enquête, l'élaboration d'un système intégré de traitement des données (ISSA) et un changement dans la philosophie du traitement des enquêtes.

2. Vers la micro-informatique autonome

Les problèmes de la grande informatique en pays en voie de développement peuvent être particulièrement contrariants pour les chercheurs spécialisés en sciences sociales. La fréquence des pannes dues à un matériel désuet, le manque de logiciels appropriés ou la pénurie de programmeurs compétents font perdre du temps aux utilisateurs; les chercheurs sont souvent relégués à la fin d'une longue file d'attente pour le traitement.

Le temps machine et de programmation peut être difficile à obtenir parce que les services administratifs et financiers du gouvernement ont également besoin de l'ordinateur.

Équiper un organisme d'enquête de deux ou trois micro-ordinateurs, de périphériques tels que des imprimantes, de boîtes de Bernoulli et de régulateurs de voltage et de logiciels faciles à utiliser règle le problème de la disponibilité de l'ordinateur pour le traitement d'enquêtes. Les possibilités de production de rapports sont également rehaussées grâce à des logiciels graphiques et au traitement de texte. Le coût du matériel, des logiciels et des autres accessoires peut être inférieur à celui du traitement sur unité centrale dans certains pays.

Le PEDS a eu très peu de problèmes de matériel pendant ou après le traitement d'enquêtes. On achetait normalement le matériel de trois à six mois avant le travail sur le terrain et on l'utilisait pendant toute cette période dans les bureaux du PEDS, où firent surface les problèmes, tels que ceux encourus avec les boîtes de Bernoulli, les disques rigides et les claviers; la réparation ou le remplacement fut facile à réaliser.

L'organisation matérielle pour les premières enquêtes du PEDS en 1986 était constituée de portatifs IBM-XT ou COMPAQ, d'imprimantes de qualité et d'un régulateur de courant. Comme les coûts de matériel diminuaient et que la vitesse du traitement augmentait, on se procura des AT, des micros 286-12Mhz et 386-12Mhz. L'organisation matérielle actuelle pour chaque enquête consiste en un ordinateur 386 et deux ordinateurs 286, une boîte de Bernoulli, des imprimantes et des UPS ou régulateurs de voltage. Tous les micros sont équipés de disques rigides de 30 à 70 Mbytes.

Bien que le milieu physique des micros exige beaucoup moins de soins que celui d'une unité centrale, le PEDS veille à ce que la salle destinée à l'informatique soit maintenue propre et sèche, car la poussière et l'humidité sont les facteurs qui entraînent le plus de problèmes de disquette et de disque rigide. La réglementation sur le manger, le boire, la consommation de tabac et le nettoyage contribue grandement au bon fonctionnement de l'équipement.

Dans certains pays, les organismes d'enquête ont pu bénéficier des services de programmeurs en micro-informatique. Le PEDS forme néanmoins le personnel d'enquête à l'entretien et à l'utilisation de l'équipement et des autres accessoires. On a mis l'accent pour la formation, sur l'utilisation des logiciels nécessaires au traitement de l'enquête, tels que DOS et des logiciels de saisie, de vérification et de tabulation. Certaines parties du DHS Data Processing Manual complètent les manuels d'utilisation propres à chaque pièce d'équipement et logiciel.

Le PEDS a fourni à chaque organisme réalisant une enquête le logiciel SPSS-PC+ pour l'analyse des données, un progiciel de traitement de texte, tel que Volkswriter ou Word Perfect, et des logiciels utilitaires, tels que Norton's Utilities et PC Tools. Plus récemment, on a installé également des progiciels graphiques afin de rehausser les possibilités de diffusion des résultats. Malheureusement, le temps a manqué pour bien former les membres du personnel à l'usage de ces progiciels. Toutefois, ceux-ci, mus par le désir de se perfectionner, ont fini par recourir à l'auto-apprentissage.

Il est clair qu'une formation supplémentaire est nécessaire lorsque le traitement d'enquête passe à la micro-informatique. Le PEDS a créé ses propres menus pour venir en aide au débutant en informatique devant accéder à DOS et aux progiciels graphiques, utilitaires de traitement des données et de traitement de texte. Afin de simplifier le traitement, le PEDS a choisi de mettre au point un progiciel intégré pouvant réaliser la saisie des données, la vérification, l'imputation, la tabulation et la construction de fichiers.

3. La mise au point d'un système intégré d'analyse d'enquête (ISSA)

Pour les enquêtes qu'il a lancées au début de 1986, le PEDS a utilisé une combinaison de progiciels: Entry Point pour la saisie des données, CONCOR pour la vérification, et SPSS-PC pour la tabulation. Au fur et à mesure qu'il évoluait, l'ISSA fut également utilisé pour une partie de la vérification et des tabulations lors de ces premiers sondages. L'utilisation de ces divers progiciels avec des exigences pour le format de fichier et des langages de commande différents accentua le besoin d'un système intégré pouvant réaliser la saisie des données, la vérification, la tabulation et la construction de fichiers.

Le composant intelligent de saisie des données de l'ISSA a été conçu de façon à ce que l'utilisateur puisse produire un fichier de données ne contenant pas d'erreurs de structure, d'intervalle, de routine: "passez à" ou de cohérence. Une telle conception est possible grâce à la structure des données de l'ISSA, qui permet d'avoir accès à un cas pendant la saisie des données afin d'y effectuer des changements et protège le chemin des réponses. L'opérateur ne peut passer outre les intervalles ou les instructions "passez à".

Sans ces contrôles, le traitement d'une enquête avec "un logiciel intelligent de saisie de données" sur micro pourrait exiger la même vérification secondaire par lots que les méthodes traditionnelles de traitement. Des erreurs de structure, d'intervalle, de routine et de cohérence pourraient exister après la saisie des données. Les vérificateurs auraient à retourner constamment aux questionnaires, à la recherche de corrections à introduire dans le fichier de données.

Si on avait utilisé des progiciels différents pour la saisie et la vérification des données, le travail aurait été plus compliqué car il aurait fallu écrire les mêmes modications dans deux langages de programmation différents. Bien qu'il y aurait eu moins d'erreurs à corriger que dans un environnement où la saisie des données est faite de façon productive (c-à-d orienté vers la saisie rapide des données) il aurait fallu faire la vérification, et le temps dédié à la programmation et à la vérification aurait été considérable.

L'utilisation combinée de Entry Point et de CONCOR par le PEDS pour la saisie et la vérification en 1986 est un bon exemple de ce type de problème. Bien que Entry Point permettait d'effectuer des modifications d'intervalle et de routine au cours de la saisie, l'opérateur qui entre les données pouvait ne pas tenir compte de cette option. Les erreurs d'intervalle étaient signalées dans le fichier de données, mais l'option de Entry Point permettant de localiser les champs signalés ne fut pas toujours utilisée. Des valeurs hors de portée et des chemins incohérents de réponses apparaissaient donc dans les réponses et il fallut une vérification supplémentaire avec CONCOR pour les éliminer.

La structure des fichiers de données et les sauts se trouvant dans des champs ayant déjà été affichés ne pouvaient être bien contrôlés: les opérateurs de saisie ne pouvaient pas retourner aux champs affichés antérieurement afin de vérifier si les données qui y avaient été introduites étaient erronées. Par exemple, un champ contenant le nombre d'enfants nés vivants d'une interviewée ne pouvait être utilisé pour contrôler le nombre d'entrées pour les naissances car il était impossible de retourner au champ si les données avaient été mal saisies. Il fallut inclure toutes les modifications de structure dans un programme secondaire de vérification CONCOR et retourner au fichier de données Entry Point pour effectuer les corrections.

Les programmes de saisie de données permettant à l'opérateur de passer outre les corrections à effectuer signalent souvent les champs qui contiennent des erreurs. Comme

il faut mettre un indicateur en mémoire dans chaque champ, la mise en mémoire des données peut se faire en un format propre au progiciel utilisé; avant que les données puissent être utilisées avec d'autres progiciels, elles devront être transcrites en un ensemble de données en code ASCII. Une des tâches les plus longues du traitement des enquêtes en 1986 fut la transcription de Entry Point à ASCII pour le traitement avec CONCOR. Le PEDS rejeta cette option de format pour les fichiers ISSA; les fichiers de données créés par ISSA sont en format ASCII.

Beaucoup de progiciels de saisie des données traitent les données de façon séquentielle; afin de recouvrer le 500e cas, il faut lire les 499 cas qui le précèdent. ISSA produit un fichier index pour chaque fichier de données créé lors de la saisie. S'il est détruit, l'index peut être reconstitué grâce à un programme utilitaire ISSA. L'index permet le retrait immédiat d'un cas pour sa mise à jour à même ISSA.

Un dictionnaire ISSA, qui définit le fichier de données pour la saisie, peut être utilisé pour les composants en lots du système: la vérification, l'imputation, la tabulation et la construction de fichiers. Le dictionnaire comprend les intervalles des valeurs de données, les noms des valeurs et des variables. Sa construction et son emploi sont complètement commandés par menu; il comprend un éditeur pour la conception de masques de saisie.

Le composant de tabulation de l'ISSA fut mis au point afin de produire tous les tableaux requis pour les rapports d'enquête du PEDS. L'ISSA peut produire des tabulations recoupées ou des moyennes simples, des tabulations recoupées ou des moyennes composées constituées de variables en colonnes et(ou) rangées multiples, et des tableaux constitués à partir de plusieurs manipulations de tableaux. Afin d'accélérer le traitement des lots, l'ISSA peut transposer un fichier de données de cas en un fichier de données de variables. Comme test d'évaluation des performances, on a demandé au système de produire plusieurs tabulations recoupées à partir d'un fichier comportant plus de 6,000 cas sur un micro 386; l'épreuve fut réussie en environ une minute.

Le questionnaire de PEDS produit un fichier à deux niveaux, c'est-à-dire contenant des données sur les ménages et les individus. Puisque beaucoup de chercheurs voudront sans doute avoir accès à SPSS/PC+ pour l'analyse statistique, l'ISSA offre un programme utilitaire pouvant produire une table à deux dimensions pouvant être lue par SPSS/PC+. Un fichier de commandes SPSS/PC+ servant à lire le fichier et poser des noms sur les variables et les valeurs est également produit.

En plus des données récoltées aux niveaux des ménages et des individus, le PEDS a récolté des données au niveau de l'unité primaire d'échantillonnage (UPE). Afin de pouvoir incorporer les données du niveau communautaire au fichier PEDS standard de données sur les ménages et les individus, on a conçu l'ISSA de façon à ce qu'il puisse traiter plusieurs fichiers à la fois. Jusqu'à trois fichiers avec index peuvent être lus et mis à jour pendant le traitement d'un fichier principal de saisie.

Les besoins du PEDS exigèrent la mise au point d'un puissant progiciel de saisie des données, de vérification et de tabulation. À la fin de 1986, l'ISSA satisfaisait presque tous les besoins de traitement des données du PEDS. L'évolution de l'ISSA au cours de la Phase 2 du PEDS comprendra une tabulation plus facile à utiliser et la possibilité de calculer les erreurs d'échantillonnage.

4. Une nouvelle philosophie du traitement d'enquêtes

Il est évident que la disponibilité des micro-ordinateurs et la mise au point de logiciels puissants ont grandement contribué au succès obtenu par le PEDS; un tel succès n'aurait cependant pas été atteint s'il n'y avait eu un changement dans la philosophie du traitement d'enquêtes. Les micros et les logiciels auraient pu être utilisés pour continuer d'effectuer le traitement de la façon traditionnelle illustrée à la figure 1, dépendante de la vérification secondaire par lots pour l'obtention d'un fichier sans erreurs au cours de la saisie. Le PEDS a néanmoins choisi d'adopter une approche différente, telle qu'illustrée à la figure 2, mettant l'emphase sur l'obtention d'un fichier sans erreurs au cours de la saisie.

Sachant que le traitement des enquêtes serait centré sur des méthodes intelligentes de saisie des données, le PEDS a pu mettre au point un questionnaire auto-codé. En disposant soigneusement les réponses aux questions et les casiers pour le codage des questions ouvertes, le traitement PEDS n'a pas eu à effectuer la transcription des données, qui est longue et sujette à erreur. Les opérateurs de saisie peuvent facilement suivre la suite des réponses sur le questionnaire.

Même pour les premières enquêtes en 1986, on a dû soigneusement choisir les agents de saisie des données car cette opération comprenait également de la vérification lorsque des problèmes surgissaient. Dans presque tous les cas, les agents de saisie ont reçu la même formation que les interviewers car ils devaient bien comprendre le questionnaire utilisé.

Puisque la saisie de données faite par le PEDS comprend également la vérification assistée par ordinateur, très peu de temps est dédié à la vérification préliminaire au bureau. Lorsque les questionnaires reviennent du terrain, on s'assure qu'il n'en manque aucun et on vérifie seulement l'ordre des entrées sur les naissances, la contraception et la santé. Le reste de la vérification est effectué au cours de la saisie, sauf la vérification de la cohérence des cheminements internes complexes.

Les programmes de saisie de données utilisés pour les premières enquêtes ne pouvaient pas contrôler la structure, les intervalles, les sauts et la cohérence aussi rigoureusement que ceux des enquêtes actuellement en cours. En fait, même les premières enquêtes entièrement traitées avec l'ISSA n'exploitèrent pas complètement le potentiel du système; ni la structure ni la cohérence interne ne furent complètement contrôlées.

Actuellement, tous les contrôles de structure, d'intervalles et de routine sont programmés pour la saisie. La majorité des modifications de cohérence sont réalisées par les opérateurs de saisie, à moins que l'équipe d'enquête considère que les opérateurs de saisie ne sont pas assez compétents pour effectuer une telle tâche. Le cas échéant, les modifications de cohérence ne sont indiquées que sous forme d'avertissement, ne permettant pas la correction, mais laissant à l'opérateur le choix de corriger des valeurs mal introduites.

La vérification secondaire des données visant à corriger l'incohérence n'ayant pas été éliminée au cours de la saisie se fait normalement dans les deux semaines après la saisie des UPE. Il a fallu beaucoup moins de temps pour repérer les questionnaires à problème, ceux-ci se chiffrant à quelques centaines plutôt qu'à quelques milliers. En moyenne, moins de deux passages en machine ont été nécessaires au cours de la vérification des ensembles de données des UPE. Comme l'ISSA empêche l'introduction d'erreurs d'intervalles, de routine et de structure au cours de la mise à jour, il est rare que

les corrections entraînent de nouvelles erreurs. Le PEDS a pu gagner un temps considérable en utilisant cette méthode plutôt qu'en effectuant les nombreux passages en machine requis par la vérification selon le traitement traditionnel.

L'entrée des données du PEDS commence peu après le début du travail sur le terrain -normalement après deux ou trois semaines. En commençant aussi tôt, on peut rapporter tout défaut éventuel du questionnaire à l'équipe sur le terrain, permettant à celle-ci d'améliorer son rendement.

Les données ont été entrées sur de petits fichiers de données: un par UPE. On évite ainsi tout désastre irréparable causé par un défaut du matériel ou une erreur de l'opérateur. Les données ont normalement été saisies sur disquettes souples; on a cependant aussi entré des données sur disque rigide ou sur des cartouches de Bernoulli.

Les opérateurs durent recevoir une formation supplémentaire considérable, en plus de celle qu'ils ont reçue en même temps que les interviewers. Le DHS Data Processing Manual donne des lignes directrices sur l'entrée et la vérification de données. Les instructions contenues dans le DHS Data Processing Manual pour la saisie et la vérification ont été révisées en collaboration avec le personnel d'enquête avant qu'on ne s'en serve et toute adaptation des instructions entraînée par l'application spécifique d'une enquête dans un pays donné a été documentée. Ensuite, des nouvelles lignes directrices ont été révisées par le personnel de traitement afin d'assurer l'uniformité des corrections.

On a demandé au personnel de traitement de conserver les listes des détails de la saisie et des erreurs de questionnaire rencontrées lors de la saisie et de la vérification secondaire. Ces listes documentent les problèmes de saisie et de vérification qui se sont déclarés lors d'enquêtes spécifiques.

Le choix et la formation d'opérateurs de saisie et de superviseurs qualifiés ont été déterminants pour le succès du PEDS. La nécessité de prendre des décisions et de réaliser des corrections au cours de la saisie allonge considérablement cette opération, mais le temps global requis pour le traitement de l'enquête s'en trouve sensiblement réduit. Des études préliminaires sur la qualité des données ont mené à des conclusions positives.

5. Les interviewers assistés par ordinateur

Le PEDS a démontré qu'en effectuant une saisie centralisée des données, on peut obtenir en une semaine ou deux des données d'enquête prêtes à être analysées. On s'intéresse néanmoins de plus en plus aux interviewers assistés par ordinateur, non pas tant comme moyen de gagner du temps mais plutôt pour améliorer la qualité des données recueillies.

À la fin de 1987, le PEDS mena un essai sur le terrain au Guatemala avec des ordinateurs portatifs. Le texte entier du questionnaire fut transcrit sur des écrans de saisie de données. On a formé les interviewers pour qu'ils puissent utiliser l'ISSA pour introduire les réponses à leurs questions. On ne rapporta aucun problème majeur; en fait, les commentaires des interviewers étaient extrêmement positifs.

Une autre étude menée par des interviewers assistés par ordinateur est prévue pour la Phase 2 de PEDS. Les objectifs de cette étude n'ont pas encore été définis. On tentera de voir par cette étude si la qualité de l'information peut être rehaussée en résolvant les problèmes d'incohérence au cours de l'entrevue.

6. Conclusion

Les micro-ordinateurs se sont avérés d'excellents instruments pour le traitement des enquêtes, surtout lorsqu'ils sont utilisés en conjugaison avec un solide système de traitement comme l'ISSA. Le PEDS pour la Colombie a terminé le travail sur le terrain en décembre 1986 et le Regional Population Center (CCRP) présentait les résultats préliminaires lors d'un colloque en janvier 1987. Le Department of Census and Statistics au Sri Lanka présentait les résultats finaux du PEDS lors d'un colloque en novembre 1987 pour le travail réalisé sur le terrain au printemps de 1987. Le rapport préliminaire du PEDS pour le Ghana fut publié en août 1988, seulement deux mois après que le Chana Statistical Service ait terminé le travail sur le terrain.

Le programme d'enquêtes démographiques et relatives à la santé a réalisé avec succès des enquêtes de la sorte dans de nombreux pays en voie de développement.

Figure 1 PRÉPARATION ET SAISIE TRADITIONNELLES DES DONNÉES

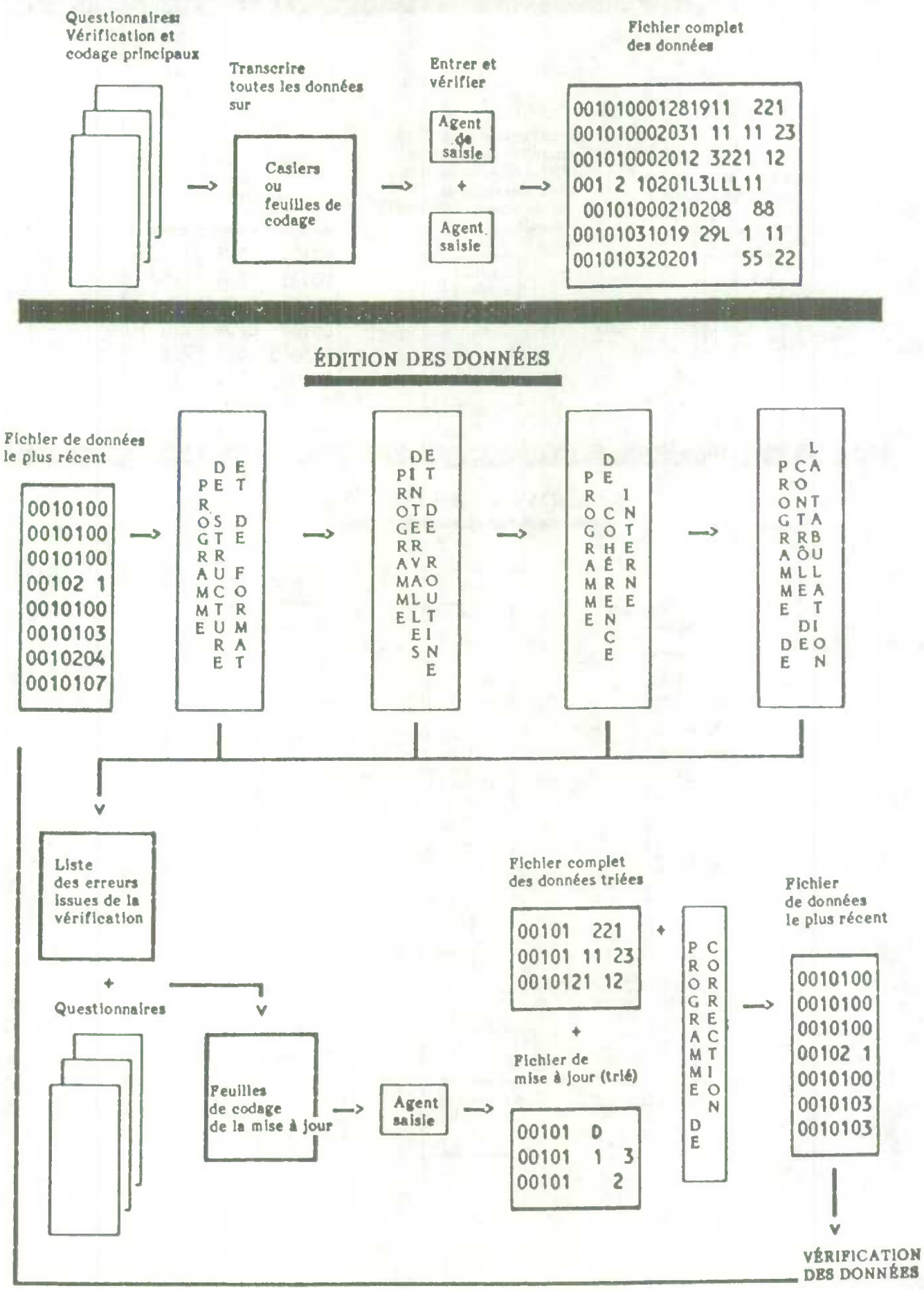
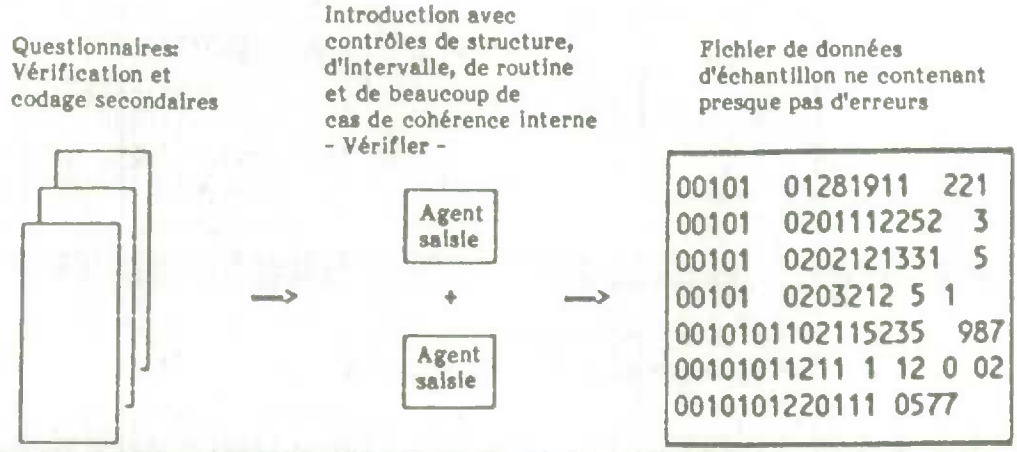


Figure 2

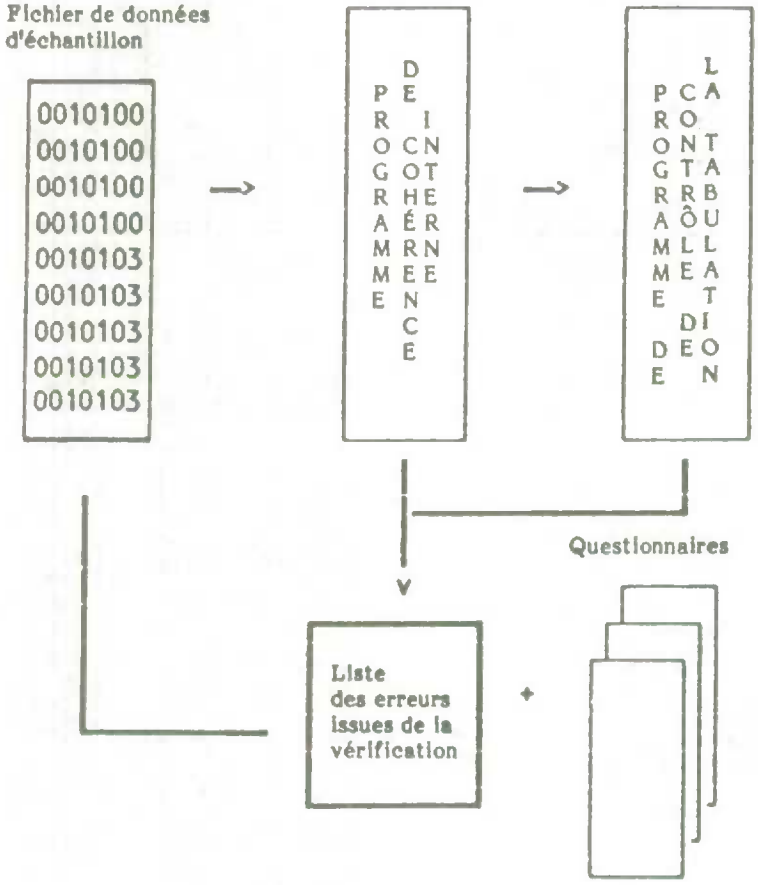
PRÉPARATION ET SAISIE DES DONNÉES DU PEDS

PEDS:



VÉRIFICATION DES DONNÉES

PEDS:



UNE COMPARAISON DE PLUSIEURS LOGICIELS POUR VÉRIFIER LES DONNÉES D'ENQUÊTE ET DE RECENSEMENT

P. COTTON,

Consultant en Statistique

RÉSUMÉ

Le présent document présente un ensemble de critères pouvant servir à évaluer les capacités des progiciels de contrôle des données d'enquête et de recensement. Ces critères portent notamment sur les fonctions de base d'un programme de contrôle, la stratégie de contrôle, la transférabilité, le niveau d'intégration, le mode d'exploitation et la documentation d'accompagnement. En outre, nous examinons certaines fonctions avancées, notamment un dictionnaire intégré de données, un programme d'entrée de données et des capacités de tabulation. À partir de ces critères, nous comparons ensuite quatre nouveaux progiciels de contrôle exploitables sur ordinateur personnel IBM.

1. Introduction

Pour retirer des résultats précis d'une enquête ou d'un recensement, un concepteur d'enquête doit s'efforcer d'éviter, ou du moins de minimiser, les erreurs au niveau des données recueillies. Bien que toutes les sources d'erreur ne puissent malheureusement pas être éliminées, la détection et la correction des erreurs constituent néanmoins des étapes importantes de l'exécution d'une enquête.

Avant l'apparition des ordinateurs de la première génération, l'essentiel des opérations de détection et de correction des erreurs de données d'enquête se faisait manuellement, soit sur le terrain, soit au siège de l'organisme responsable de l'enquête, ce qui exigeait une main d'oeuvre de bureau très nombreuse. Avec l'arrivée des mini-ordinateurs et d'ordinateurs centraux très puissants, de nombreux organismes de statistique ont commencé à automatiser la détection de certaines erreurs au moyen de programmes informatiques personnalisés, l'intervention humaine servant à décider des mesures correctives à prendre.

Au cours des années 1970, on a vu apparaître des systèmes informatiques personnalisés capables de détecter et de corriger automatiquement certains types d'erreurs de données d'enquête. Vu le degré élevé de sophistication et les coûts d'entretien de tels systèmes, les organismes de statistique se sont mis à investir dans l'élaboration de progiciels généralisés de contrôle informatique pouvant servir à différentes enquêtes (Graves, 1976). Bien que plusieurs de ces progiciels généralisés ou leurs descendants directs soient encore utilisés aujourd'hui, une nouvelle génération est disponible depuis peu. Ces progiciels qui sont exploitables sur des ordinateurs personnels IBM que l'on trouve désormais partout, offrent au concepteur d'enquête la possibilité de choisir, quand il existe, le progiciel le plus approprié aux besoins de son enquête. La présente communication vise à aider le concepteur à évaluer celui qui convient le mieux à telle ou telle enquête ou recensement.

Le Chapitre 2 décrit un ensemble de critères permettant au concepteur d'enquête d'évaluer un progiciel de contrôle spécifique. Le Chapitre 3 reprend ces critères pour établir une comparaison entre quatre progiciels de contrôle, exploitables sur ordinateur personnel IBM seulement, à savoir :

1. Système généralisé de contrôle et d'imputation (SGCI),
2. Système intégré d'analyse des enquêtes (ISSA),
3. PCEDIT, et
4. Système intégré de traitement sur micro-ordinateur (IMPS).

Le Chapitre 4 présente un résumé des résultats et conclusions de notre évaluation.

2. Critères d'évaluation des progiciels de contrôle des données d'enquête et de recensement

Ce chapitre présente un ensemble des critères pouvant servir à évaluer les progiciels de contrôle des données d'enquête ou de recensement. Ces critères sont repris au Chapitre 3 pour comparer quatre progiciels de contrôle récemment mis sur le marché.

2.1 Fonctions de base d'un programme de contrôle

Le progiciel de contrôle des données d'enquête et de recensement doit exécuter les fonctions de base suivantes :

Description du fichier d'entrée

Le progiciel de contrôle doit offrir la capacité de décrire la présentation du questionnaire ou des données d'enquête. Habituellement, il s'agit de décrire la présentation de chaque "enregistrement" du fichier des données d'enquête ainsi que les "zones" de chaque enregistrement. La description doit permettre de spécifier une zone determinant le type d'enregistrement, ainsi que des zones "clés" identifiant de façon unique chaque questionnaire. Le progiciel de contrôle permet parfois à l'utilisateur de spécifier comment les zones clés correspondent à des strates différentes de l'enquête (régions géographiques, par exemple).

Contrôles structurels

Le progiciel de contrôle doit permettre d'établir la présence de tous les enregistrements d'un questionnaire et l'absence de tout enregistrement superflu. Par exemple, le programme de contrôle des données d'un recensement démographique doit permettre de déterminer quels sont les enregistrements relatifs à tel ménage et à telle personne qui correspondent à tel questionnaire.

Contrôles de la validité des valeurs

Le progiciel de contrôle doit déterminer si la valeur d'une zone se situe à l'intérieur ou à l'extérieur des limites de validité de cette zone. Ces contrôles ne s'appliquent qu'à des zones ou variables uniques.

Contrôles de cohérence

Le progiciel de contrôle doit pouvoir comparer au moins deux zones d'un questionnaire et vérifier leur cohérence. Les zones en question peuvent faire partie du même enregistrement ou d'enregistrements différents du questionnaire.

Analyse des règles de contrôle

Pour certains progiciels de contrôle, on peut introduire un ensemble de règles de contrôle qui servent à effectuer les contrôles de validité et de cohérence d'un questionnaire d'enquête. Ces progiciels incluent habituellement un programme qui sert à déterminer si les contrôles spécifiés sont redondants ou contradictoires, et si l'on peut dériver des contrôles supplémentaires non spécifiés par l'utilisateur à partir des contrôles initiaux.

Correction automatique

Le progiciel de contrôle doit avoir la capacité de corriger les erreurs détectées par les contrôles de validité ou de cohérence. Les corrections peuvent être apportées directement par le programme de contrôle de l'utilisateur, utilisant des techniques telles que le "hot deck" ou le "cold deck", ou encore automatiquement par le progiciel de contrôle lui-même. Quand un programme de contrôle effectue automatiquement les corrections, le principe consiste habituellement à modifier le moins de zones possible.

Fichier de données corrigées

Le progiciel de contrôle doit créer un fichier de données corrigées. Il peut s'agir d'un nouveau fichier de format identique au fichier d'entrée, ou bien les modifications peuvent être apportées directement au fichier d'entrée d'origine.

Fichier de données extraites

Le progiciel de contrôle doit permettre à l'utilisateur de créer un fichier contenant n'importe quelle combinaison de données du fichier d'entrée d'origine, et n'importe quelles variables dérivées par le programme de contrôle. L'utilisateur doit pouvoir spécifier le format de ce fichier, c'est-à-dire données ASCII, format SPSS/PC, etc.

Rapports

Le progiciel de contrôle doit pouvoir générer des rapports sur les erreurs détectées et sur les corrections. Ces rapports peuvent inclure certains ou l'ensemble des éléments suivants :

1. Pour chaque règle de contrôle, le nombre de fois où elle a détecté une erreur
2. Le nombre de fois où des valeurs particulières ont été affectées à une variable
3. Pour chaque questionnaire, les règles de contrôle qui ont détecté des erreurs.

L'utilisateur doit pouvoir choisir les rapports de contrôle qu'il veut générer et les niveaux géographiques auxquels il veut qu'ils soient générés. Il doit pouvoir demander le vidage de certaines ou de toutes les zones d'un questionnaire.

2.2 Fonctions avancées d'un programme de contrôle

Plusieurs des programmes généralisés de contrôle disponibles depuis un certain temps et exploitables sur mini-ordinateurs et ordinateurs principaux seulement, exécutent les fonctions de base que nous venons de décrire. La nouvelle génération des progiciels intégrés de contrôle exploitables sur ordinateur personnel IBM combinent ces fonctions de base à l'ensemble des fonctions avancées suivantes :

Programme de dictionnaire de données

Le dictionnaire de données est le "ciment" qui soude un progiciel intégré de contrôle. Le dictionnaire de données contient les méta-données (c'est-à-dire les données sur les données) nécessaires au concepteur d'enquête. Avec un progiciel bien intégré, ces méta-données n'ont à être décrites qu'une seule fois par le concepteur. Les mêmes méta-données servent ensuite à chacun des programmes constitutifs du progiciel intégré. Ce type d'environnement est préférable, dans la mesure où l'utilisateur n'est pas obligé de décrire séparément les données d'enquête pour chaque programme.

Le dictionnaire de données contient des renseignements sur la présentation des données d'enquête. Chacune des variables de l'enquête ou des variables dérivées est affectée d'un nom logique auquel peut se référer chacun des programmes constitutifs du progiciel intégré. Pour chaque zone, le dictionnaire de données comprendra la définition d'une valeur manquante de même que la définition d'une valeur non applicable. Le dictionnaire devra également permettre à l'utilisateur de définir des noms ou labels descriptifs correspondant aux valeurs des variables encodées, et par la suite de se référer à ces noms plutôt qu'aux codes réels employés dans les programmes intégrés. Ces types de méta-données assurent à l'utilisateur un niveau d'indépendance des données qui protège son investissement contre toute modification légère des formats de ses données.

Le programme de dictionnaire de données doit offrir la capacité de créer un fichier de sortie englobant toutes les méta-données de l'enquête et pouvant être utilisé comme document d'accompagnement de l'enquête.

Programme d'entrée des données

Pour que l'utilisateur puisse introduire interactivement ses données d'enquête sur son ordinateur personnel IBM, le progiciel intégré de contrôle doit être doté d'un programme

d'entrée des données. Celui-ci doit être bien intégré au dictionnaire de données du progiciel et si c'est possible, aux modules de base de détection et de correction des erreurs du programme de contrôle. Les exigences d'un programme d'entrée des données d'enquête et de recensement peuvent se résumer par la liste de caractéristiques essentielles suivantes (Diskin, 1986):

1. écrans programmables d'entrée des données
2. capacité de détecter les erreurs pendant l'entrée des données, et notamment par des contrôles de validité et de cohérence intra-zone
3. définition du schéma des instructions de "passer à..." et affectation de la valeur non-applicable aux zones évitées
4. duplication et augmentation automatiques du nombre des zones
5. mise en séquence programmable de l'entrée des zones (commande du curseur)
6. capacité de verification des données
7. extraction et modification d'un questionnaire ou d'un enregistrement
8. vitesse d'exécution
9. statistiques reliées à l'opérateur.

Programme d'extraction ou de tabulation

Quand un progiciel intégré de contrôle comprend un programme de tabulation, le concepteur d'enquête peut l'utiliser avant, pendant et après le contrôle des données d'enquête. Sinon, le concepteur d'enquête doit recourir à un programme séparé, ce qui l'oblige à utiliser un autre langage de description des données et un autre langage de manipulation des données.

2.3 Intégration

Parmi les micro-ordinateurs que nous trouvons aujourd'hui sur le marché, nous sommes souvent forcés de choisir entre des produits ne possèdant pas les mêmes niveaux d'intégration. Par exemple, nous pouvons acheter soit un très bon traitement de texte, soit un très bon programme de chiffrier, soit un très bon programme de graphisme, ou peut-être même un seul progiciel intégrant ces trois fonctions à un coût total beaucoup moins élevé, mais chacune étant de qualité moyenne.

Un progiciel bien intégré se caractérise par un interface utilisateur unique et cohérent pour toutes les fonctions, par exemple Lotus 1-2-3. Par exemple, les menus auront fondamentalement la même apparence pour chaque programme et les mêmes touches permettront d'exécuter les mêmes fonctions. Un progiciel de contrôle moins bien intégré exige plusieurs interfaces utilisateur; il faudra plus de temps pour apprendre à se servir de la machine et elle ne sera pas aussi facile à utiliser.

2.4 Stratégie de contrôle

Dans la conception de toute enquête, l'essentiel consiste à décider du moment où la détection et la correction des erreurs des données d'enquête doivent intervenir. Pour des enquêtes de moindre envergure, il peut encore être possible d'effectuer des contrôles manuels des questionnaires sur le terrain, ou encore après qu'ils aient été renvoyés à l'organisme responsable. Comme nous l'avons vu, c'est une option coûteuse qui n'est pas à retenir pour des enquêtes ou recensements de plus grande envergure.

Le concepteur d'enquête doit ensuite décider du degré de contrôle des données à exercer pendant et après l'entrée des données. Malheureusement, cette décision ne peut être prise isolément, dans la mesure où le degré de contrôle exercé pendant l'entrée peut avoir un impact majeur sur la durée totale de l'opération d'entrée des données.

Pour un recensement démographique, par exemple, le programme d'entrée des données n'effectue habituellement que les contrôles de validité des valeurs de zones individuelles de données, puisque tout contrôle plus systématique (c'est-à-dire contrôles de cohérence ou

intra-zone) obligerait les opérateurs chargés de l'entrée des données à corriger des erreurs plus complexes, ce qui pourrait avoir un impact majeur sur la durée totale du temps écoulé avant que les données d'enquête ne soient prêtes pour l'étape suivante du traitement. D'autre part, pour une enquête de moindre envergure mais faisant intervenir un questionnaire important et complexe (c'est-à-dire avec de nombreuses instructions de "passer à..."), le concepteur peut choisir d'utiliser un programme d'entrée des données "intelligent", c'est-à-dire capable de vérifier à l'entrée des données que chaque questionnaire a été correctement complété par l'interviewer. Cependant, pour que le concepteur d'enquête puisse choisir de ralentir l'entrée des données et vérifier que chaque questionnaire présente la structure qui convient, le volume des données doit être suffisamment faible. C'est un avantage car il est très difficile pour un programme de contrôle de corriger les erreurs d'un questionnaire aux étapes ultérieures du traitement.

Le progiciel de contrôle idéal doit permettre au concepteur d'enquête de sélectionner le degré de contrôle à exercer pendant et après l'entrée des données.

2.5 Mode d'exploitation interactif et mode d'exploitation par lots

La plupart des premiers systèmes généralisés de contrôle faisaient du traitement par lots, ayant été conçus pour de gros ordinateurs particulièrement propices à ce type de traitement. L'utilisateur devait habituellement coder ses règles ou programmes de contrôle d'entrée selon un format fixe, et faire analyser ensuite les résultats du traitement par lots par le système de contrôle. Même avec un programme syntaxiquement correct, l'analyse du travail exécuté en traitement par lots prenait beaucoup de temps, en plus d'être ennuyeuse. Le traitement par lots n'était plus un inconvénient, à partir du moment où l'utilisateur disposait d'un programme "correct" et commençait vraiment à contrôler ses données.

Aujourd'hui, grâce à la puissance relative et au coût peu élevé des ordinateurs personnels, les systèmes interactifs sont devenus monnaie courante. L'utilisateur d'un ordinateur personnel s'attend à ce que le progiciel puisse contrôler la syntaxe de ses données au moment de l'entrée, et à pouvoir tester son "programme" à n'importe quelle étape de son développement.

Vu le coût élevé de la main d'oeuvre humaine et celui toujours décroissant des ordinateurs, les capacités interactives d'un progiciel intégré de contrôle revêtent une importance majeure. Par ailleurs, les concepteurs d'une enquête ou d'un recensement de très grande envergure souhaitent encore pouvoir recourir au traitement par lots. C'est pourquoi un progiciel de contrôle fonctionnant à la fois en mode interactif et en mode de traitement par lots répondra aux besoins d'un plus grand nombre d'enquêtes.

2.6 Transférabilité

Bien que les micro-ordinateurs d'aujourd'hui soient très puissants, il existe encore certaines tâches de contrôle complexes qu'il vaut mieux exécuter sur un ordinateur central ou mini-ordinateur plus important et plus puissant. Un progiciel de contrôle transférable, pouvant être utilisé à la fois sur des micro-ordinateurs, sur des mini-ordinateurs et sur des gros ordinateurs centraux permettra à l'utilisateur d'élaborer une application expérimentale sur un ordinateur personnel, et de transférer directement le programme et/ou les données de contrôle sur un système beaucoup plus puissant, pour les phases d'exécution. Étant donné qu'apprendre à se servir de plusieurs ordinateurs différents coûte très cher, un progiciel de contrôle dont l'interface utilisateur est transférable sur des systèmes de taille différente simplifiera beaucoup l'apprentissage.

Le monde des ordinateurs connaît aujourd'hui une évolution rapide et, par conséquent, la transférabilité des progiciels sur les systèmes d'exploitation de micro-ordinateurs comme le OS/2, le Microsoft Windows et le UNIX peut également s'avérer un facteur important pour certaines applications spécifiques.

2.7 Structures complexes de données

Dans certains cas, la capacité d'un progiciel de contrôle de manipuler facilement des structures complexes de données est un facteur décisif. Par exemple, le concepteur d'une enquête sur les ménages qui a besoin de renseignements à la fois au niveau des ménages et des individus recherchera un progiciel de contrôle capable de traiter des données hiérarchiques. Certains questionnaires comportent une série de questions qui se répètent un certain nombre de fois, comme par exemple les antécédents médicaux depuis la naissance, dans le cas d'une enquête sur la santé. Pour ce genre d'enquêtes, la capacité du progiciel de contrôler facilement un groupe de questions répétitives s'avérera très utile.

Un progiciel dont la composante entrée des données facilite le contrôle en direct du questionnaire sera très apprécié du concepteur d'enquête dont la tâche consiste à introduire et à contrôler les données d'un vaste questionnaire assorti d'un schéma compliqué d'instructions de "passer à...".

Si le progiciel intégré est doté d'un système de tabulation, la capacité de créer certaines structures spéciales de stockage, pouvant être optimisées pour les fonctions d'extraction, peut s'avérer un important critère de sélection.

2.8 Données qualitatives et données quantitatives

La majorité des enquêtes sur les ménages et des recensements démographiques recueillent surtout des données codées ou qualitatives, alors que la plupart des enquêtes économiques recueillent des données numériques ou quantitatives. Bien que l'on puisse utiliser des progiciels de contrôle polyvalents pour ces deux types de données, certaines techniques avancées de détection et de correction automatiques permettent de traiter plus spécifiquement soit les unes, soit les autres. Le type de données à vérifier sera souvent l'un des critères les plus importants dans le choix du progiciel de contrôle le plus approprié. De fait, le concepteur d'enquête sera parfois amené à choisir entre une application personnalisée et un progiciel généralisé spécifique.

2.9 Type d'entrée

Le type des entrées requises et, par conséquent, l'expérience que l'utilisateur devra avoir pour utiliser sans difficulté et correctement le progiciel de contrôle sont des éléments importants dans le choix d'un progiciel. Le type d'entrée permet également de déterminer si l'utilisateur du progiciel aura besoin ou non d'un apprentissage spécialisé.

Un grand nombre de progiciels de contrôle s'adressent surtout à des programmeurs, dans la mesure où les langages d'entrée sont déjà des langages de programmation spécialisés de la troisième génération. Ceux qui s'adressent à des utilisateurs ayant une expérience limitée de la programmation tolèrent un plus grand nombre de paramètres d'entrée. Par exemple, plusieurs logiciels de contrôle exigent la spécification d'un seul ensemble de règles de contrôle. Ces règles sont ensuite analysées et, le cas échéant, un programme spécial établi à partir de ces règles détecte et corrige automatiquement les erreurs des données d'enquête.

2.10 Efficience

Selon l'envergure de l'enquête et selon les ressources informatiques disponibles, l'efficience d'un progiciel de contrôle peut être un critère de sélection très important. Par exemple, si le programme d'entrée des données d'un progiciel intégré de contrôle n'admet pas des taux d'entrée de données supérieurs à 10,000 frappes à l'heure, ce progiciel sera inapproprié pour une enquête ou un recensement où l'entrée des données est habituellement l'élément critique, par rapport à l'échéancier global du projet.

Il est nécessaire d'accorder une attention particulière à l'évaluation de l'efficience des progiciels de contrôle exécutant la correction automatique des données non valides. Ces progiciels, et en particulier ceux qui corrigent des variables quantitatives, s'appuyent sur des algorithmes pouvant s'avérer très coûteux à exécuter lorsque le nombre des enregistrements, des règles de contrôle ou des variables est élevé.

Au chapitre de l'efficience, il est peut-être encore plus essentiel de savoir si le programme de contrôle de l'utilisateur créé à partir du progiciel de contrôle est interprété ou compilé. Bien que les programmes interprétés entraînent des frais généraux moins élevés lorsqu'ils sont créés, ils fonctionnent en général plus lentement que des programmes déjà compilés et n'ayant qu'à être exécutés. Un concepteur à la recherche d'un système de contrôle interactif choisira habituellement un système interprétatif, alors que celui qui recherche un système à traitement par lots pour effectuer une enquête ou un recensement de plus grande envergure préférera un système compilé. La documentation d'accompagnement du progiciel de contrôle doit indiquer sa vitesse d'exécution relative et, si possible, fournir un échantillon ordinolingue comprenant suffisamment de données pour permettre l'évaluation du progiciel dans des environnements différents.

La sélection d'un progiciel intégré de contrôle exploitable sur ordinateur personnel IBM dépend également du type d'ordinateur sur lequel il fonctionne de façon optimale. Le processus de sélection doit permettre de déterminer les éléments suivants, qui sont nécessaires au progiciel de contrôle :

1. quantité de mémoire vive (RAM)
2. quantité d'espace sur disque nécessaire pendant l'exécution pour le logiciel, les données d'enquête et les fichiers temporaires
3. présence d'un co-processeur numérique
4. besoins spéciaux, notamment "mémoire supplémentaire", "contrôle du graphique", etc.
5. niveau minimum des ressources requises par le programme d'entrée des données pour pouvoir créer des postes peu coûteux d'entrée des données.

2.11 Documentation et apprentissage

Avant de choisir un progiciel, l'utilisateur doit tenir compte de la documentation d'accompagnement ainsi que des cours offerts. S'il est important de pouvoir s'appuyer sur des documents de référence écrits, il peut s'avérer encore plus essentiel d'avoir accès en direct à des instructions et documents d'appui, surtout pour des programmes interactifs utilisables sur ordinateur personnel IBM.

La documentation d'accompagnement doit contenir les éléments suivants :

1. instructions de mise au point et de "mise en marche"
2. renseignements à l'intention des novices
3. guide de l'utilisateur et
4. manuel de référence.

La documentation d'accompagnement d'un progiciel de contrôle doit présenter plusieurs exemples différents, afin que les utilisateurs novices puissent se familiariser avec les caractéristiques du progiciel. Ces exemples doivent être rédigés sous forme ordinolingue et, si possible, se retrouver dans la documentation écrite. Le guide de l'utilisateur doit également inclure des instructions sur la façon d'utiliser le progiciel pour résoudre différents types de problèmes.

2.12 Coût et valeur

Pour les concepteurs d'enquête disposant d'un budget limité, le coût du progiciel constitue un important critère de sélection. Ce coût inclut le coût du progiciel de contrôle lui-même et celui de tout autre logiciel spécialisé nécessaire à son utilisation. En effet, l'utilisateur devra peut-être acheter et apprendre à utiliser un logiciel de base de données afin de pouvoir créer les structures sous-jacentes nécessaires à l'établissement des fichiers de ses données d'enquête. Il ne faut pas uniquement évaluer le coût en dollars, mais tenir compte également de l'intérêt du logiciel acheté. Un logiciel autonome n'ayant pas la fonctionnalité recherchée n'aura aucun intérêt pour un concepteur d'enquête.

3. Évaluation des progiciels de contrôle

3.1 Système généralisé de contrôle et d'imputation (SGCI)

Le système SGCI (1988) est un progiciel numérique de contrôle que Statistique Canada est en train d'élaborer pour le traitement des données d'enquêtes économiques. Le SGCI servira à imputer les erreurs restantes, une fois que le contrôle et le suivi préliminaires en auront éliminé la plupart, exception faite des erreurs mineures et des erreurs non détectées. Le SGCI est le fruit de recherches antérieures de Statistique Canada consacrées au contrôle des données numériques (Sande, 1979) et des données codées (Felligi et Holt, 1976).

La présente évaluation porte sur un exemplaire de la version 4.2 du SGCI, fourni à l'auteur par Statistique Canada aux seules fins de ladite évaluation. Le logiciel du SGCI n'est pas complet et il n'est pas distribué à l'heure actuelle à l'extérieur de Statistique Canada.

3.1.1 Description générale

Le SGCI regroupe un ensemble de modules de programmes séparés, pouvant être utilisés par les concepteurs d'enquête pour contrôler leurs données numériques. Au lieu d'élaborer un système monolithique complexe, les concepteurs du SGCI ont préféré l'approche modulaire. Ainsi, l'utilisateur du système a la possibilité de créer des prototypes de module individuels, d'améliorer certains modules sans affecter les autres et d'ajouter de nouveaux modules, au fur et à mesure qu'il se familiarise avec le système. De la même façon, les concepteurs d'enquête peuvent n'utiliser que les modules répondant aux besoins particuliers de leur enquête.

Le SGCI est intégré au système de gestion de base de données (SGBD) ORACLE (1988). Il est fondé sur l'hypothèse que les données à contrôler existent déjà dans une base de données ORACLE. Les utilisateurs du SGCI doivent donc, au minimum, apprendre le langage structuré de recherche (SQL) qu'utilise ORACLE, afin de pouvoir créer et manipuler "les tables" de données de la base de données ORACLE. Le SGBD ORACLE a été choisi pour assurer l'organisation et la manipulation des données d'enquête et des renseignements sur les commandes du SGCI, en raison de sa transférabilité sur une gamme variée d'ordinateurs. Le SGCI et ORACLE sont utilisés à l'heure actuelle par Statistique Canada sur des ordinateurs personnels IBM, sur un système UNIX à utilisateurs multiples et sur un très gros ordinateur central IBM.

On prévoit que le SGCI comprendra les six principaux modules suivants :

1. description du questionnaire
2. description et analyse des contrôles quantitatifs
3. description et analyse des contrôles statistiques
4. localisation des erreurs
5. imputation
6. rapports

Le concepteur d'enquête utilisera d'abord le module de description du questionnaire pour indiquer au SGCI quelle est ou quelles sont la(les) table(s) ORACLE qui contiendront les données d'enquête. Les données d'enquête à contrôler seront stockées dans une seule table ORACLE mais, dans le cas d'enquêtes mensuelles ou annuelles, on pourra également indiquer au SGCI quelles sont les tables qui contiennent des données historiques. Le concepteur d'enquête devra ensuite définir au moins une strate de données. Le traitement ultérieur des données d'enquête (c'est-à-dire localisation et imputation des erreurs) sera effectué par strate. Cela permettra au concepteur d'enquête d'élaborer différents ensembles de règles de contrôle pour différents sous-ensembles de données d'enquête.

Tous les renseignements fournis par le concepteur d'enquête et tous les résultats produits par les différents modules du SGCI sont également stockés dans des tables ORACLE. Ces tables sont établies au moment de la mise au point du SGCI et peuvent être examinées en tout temps par l'intermédiaire du SQL.

Le concepteur d'enquête doit ensuite spécifier les règles de contrôle à appliquer aux données d'enquête. Chaque règle de contrôle sera désignée par un nom, pour fins de référence future, et sera exprimée sous forme linéaire au moyen de constantes et de noms de "colonne" renvoyant aux tables ORACLE contenant les données d'enquête. Par exemple, pour le SGCI, les règles de contrôle de la validité des données seront les suivantes:

0.5 * VENTES <= DÉPENSES
DÉPENSES <= VENTES
BLÉ + SEIGLE + MAIS < TOTAL - JACHÈRE

Le SGCI n'oblige pas l'utilisateur à exprimer les règles de contrôle sous une forme canonique; en effet, il lui offre un programme qui détermine automatiquement la forme canonique de chacune. Une fois que le concepteur d'enquête a introduit les règles de contrôle de base, elles peuvent être regroupées en différents groupes de contrôles applicables à différentes strates des données d'enquête.

Lorsque le concepteur a défini un groupe de contrôles, l'étape suivante consiste à les analyser. Le SGCI est doté de fonctions permettant de vérifier l'exactitude des contrôles (existence de colonnes dont le nom n'est pas défini dans une table ORACLE), de générer des contrôles implicites et de produire des enregistrements fictifs ayant franchi tous les contrôles mais se situant aux extrémités de la région définie comme acceptable par les contrôles linéaires effectués par le groupe de contrôles. Ces deux derniers éléments peuvent servir au concepteur d'enquête à s'assurer du caractère approprié de l'ensemble initial des contrôles.

Le SGCI est également doté d'une fonction permettant de détecter les données extrêmes; c'est ce que l'on appelle le module de contrôle statistique. Ce module permet à l'utilisateur de déterminer les limites supérieures et inférieures d'une variable, ainsi que le ratio de la valeur d'une variable par rapport à ses valeurs antérieures. Ce module peut servir à déterminer les limites des contrôles linéaires, ou à identifier les valeurs des données extrêmes aux fins de l'imputation des erreurs.

Lorsqu'un groupe de contrôles a été défini de façon définitive, on peut l'appliquer à une strate de données d'enquête par le truchement du module de localisation des erreurs. Pour chaque enregistrement d'une strate, ce module déterminera quelles sont les règles de contrôle qui n'ont pas permis de détecter d'erreur et quel est le nombre minimum des zones à modifier pour corriger ces erreurs. Le concepteur d'enquête peut influencer le choix des zones à retenir pour l'imputation des erreurs, en appliquant des pondérations aux zones auxquelles les règles de contrôle ont été appliquées. Lorsque la localisation des erreurs est terminée, les zones retenues pour l'imputation sont consignées pour chaque enregistrement dans l'un des tableaux de commande du SGCI. Le SGCI conserve également des renseignements sur les enregistrements qui ont franchi tous les contrôles, afin de les utiliser par la suite comme donneurs à la phase d'imputation.

Le SGCI autorise trois types d'imputation. Premièrement, l'utilisateur peut spécifier un estimateur d'imputation qui déterminera la valeur corrigée d'une zone à partir d'un ensemble d'algorithmes intégrés. Ces algorithmes peuvent faire référence à la valeur historique d'une même variable ou d'une variable auxiliaire. Deuxièmement, l'utilisateur peut choisir un mode d'imputation déterministe: seront retenues pour l'imputation les zones pour lesquelles il n'existe que des solutions uniques pour chaque enregistrement d'une strate.

L'imputation par donneur constitue le troisième type d'imputation. Ce module remplace les valeurs non valides et les valeurs manquantes d'un enregistrement à partir d'un enregistrement similaire sans erreur, établi par la méthode de la recherche du voisin le plus proche. La similarité entre les enregistrements donneurs et l'enregistrement receveur présentant une erreur est estimée à partir d'un sous-ensemble composé des valeurs exactes, non manquantes. La version 4.2 du SGCI n'est pas dotée de ce module et il n'a donc pas été testé pendant l'évaluation.

Le dernier module du SGCI est le module de production de rapports. Ce module permet à l'utilisateur de générer pour un questionnaire spécifique sept rapports différents sur les contrôles effectués. Il permet, par exemple, à l'utilisateur d'imprimer tous les contrôles ou seulement un groupe, ou encore de déterminer si certains contrôles ne font partie d'aucun groupe. Le concepteur du système peut également générer plusieurs autres rapports en utilisant les capacités SQL d'ORACLE, pour imprimer le contenu de n'importe lequel des tableaux de commande du SGCI: les résultats de la localisation des erreurs peuvent être déterminés en imprimant les rangées appropriées de la table des "zones à imputer".

3.1.2 Évaluation

Fonctions de base et fonctions avancées

Le SGCI exécute directement les fonctions de base du programme de contrôle grâce à ses propres modules, et indirectement grâce à certaines capacités sous-jacentes du SGBD ORACLE. Le SGCI est fondé sur l'hypothèse que le concepteur d'enquête a constitué une table ORACLE unique pour chaque période chronologique des données à traiter. Par conséquent, le SGCI s'appuie sur les capacités de définition des données d'ORACLE pour créer la description du fichier d'entrée. Si les données d'enquête de l'utilisateur ne se trouvent pas déjà dans une table ORACLE ou si elles sont réparties entre plusieurs tables ORACLE, il faut alors utiliser le langage SQL ou une fonction utilitaire d'ORACLE pour obtenir la table unique dont le SGCI a besoin. Le SGCI n'a pas la capacité directe de décrire un questionnaire à enregistrements multiples et, par conséquent, il n'est pas conçu pour effectuer une vérification structurelle des données d'enquête de l'utilisateur.

Le SGCI présente toutes les autres caractéristiques de base d'un programme de contrôle: contrôles de validité, contrôles de cohérence, analyse des règles de contrôle, correction automatique des erreurs et établissement d'un dossier de données corrigées. Ces fonctions concernent uniquement le traitement des données quantitatives. L'utilisateur peut se servir des capacités SQL d'ORACLE pour créer un fichier de données extraites.

Les rapports produits par le SGCI sont ventilés par strate et par groupe de contrôle. Ces rapports sont les suivants : a) nombre de fois où une règle de contrôle a détecté une erreur; b) nombre de règles de contrôle qui ont détecté une erreur, par questionnaire et c) nombre total d'échecs des contrôles et nombre total des enregistrements ayant échoué au moins un contrôle. Si l'on veut obtenir d'autres rapports, on peut utiliser le SQL pour exécuter les tableaux de commande internes du SGCI. L'utilisateur peut vider un questionnaire donné au moyen de la commande ORACLE SQL SELECT.

Le SGCI dépend d'ORACLE pour l'exécution des fonctions avancées suivantes du programme de contrôle: dictionnaire de données, capacité d'entrée de données et programme d'extraction. Le dictionnaire de données ORACLE sert à définir non seulement les données

d'enquête de l'utilisateur mais aussi les données sur les commandes, lesquelles sont mises à jour par le SGCI pour l'utilisateur. Étant donné que les méta-données ORACLE sont elles-mêmes stockées sur des tables ORACLE, l'utilisateur peut manipuler ses propres données, les données sur les commandes du SGCI et les méta-données du système grâce au SQL.

Tel qu'expliqué plus haut, le mandat du SGCI n'oblige pas l'utilisateur à s'appuyer sur un programme d'entrée de données, puisque l'on suppose que le SGCI ne sert qu'à l'imputation des données n'ayant pas été corrigées lors des contrôles préliminaires exercés au moment de la collecte ou de la saisie des données. Néanmoins, si le concepteur d'enquête doit introduire des données pour tester le SGCI, il peut se servir de la commande SQL INSERT pour remplir une table de données de test. L'utilisateur qui doit faire rentrer un grand nombre d'enregistrements sur une table utilisera le programme utilitaire ORACLE approprié.

Si le langage SQL permet d'extraire et d'afficher des ensembles d'enregistrements à partir d'une ou de plusieurs tables, ses capacités de "tabulation" sont assez limitées. Certaines opérations comme la moyenne, le maximum, le minimum et la somme, sont possibles, mais il n'existe pas de fonction intégrée permettant de constituer de tables recoupées.

Intégration

Le SGCI présente un très haut niveau d'intégration puisqu'il est intégré au SGBD ORACLE. De fait, la réussite commerciale du SGBD ORACLE tient à ce qu'il a toujours employé les langages SQL*PLUS et SQL*FORMS comme interfaces utilisateurs. La version MS-DOS du SGCI évaluée par l'auteur peut être exécutée directement à partir de la ligne de commande, ou par l'intermédiaire d'un système de menus basé sur le programme SQL*FORMS. Le système de menus permet au novice de se familiariser sans difficulté avec le SGCI, tandis que l'expert peut toujours recourir au SQL s'il a besoin d'accéder directement aux tableaux de commande du SGCI.

Stratégie de contrôle et mode d'exploitation

Le SGCI obéit au principe que le contrôle des données d'enquête se fait en deux phases : la phase préliminaire englobe la saisie des données et la deuxième phase s'appuie sur le SGCI. Même si le SGCI apparaît comme un système hautement interactif lorsqu'il est exploité sur un ordinateur personnel IBM, on peut également l'utiliser directement en mode de traitement par lots grâce aux modules du programme SGCI. Cette option constitue un point fort du SGCI et elle est à retenir.

Transférabilité

L'auteur n'a évalué que la version 4.2 du SGCI, utilisée sur un ordinateur personnel IBM. Au sein de Statistique Canada, cette version est également exploitée sur un système UNIX à utilisateurs multiples et sur un ordinateur central IBM. La possibilité d'utiliser le SGCI sur ces trois architectures différentes d'ordinateur est essentiellement attribuable à la transférabilité du SGBD ORACLE.

Les modules du SGCI sont rédigés en langage de programmation transférable C et le précompilateur ORACLE PRO*C permet d'employer le SQL pour l'exécution des données d'enquête de l'utilisateur et des tableaux de commande du SGCI. Le SGCI devrait donc être transférable sur n'importe quel ordinateur muni du compilateur C approprié et pouvant recevoir le programme ORACLE.

Le système de menus basé sur le SQL*FORMS dont nous avons parlé existe pour chacun des trois ordinateurs cités. La transférabilité de l'interface utilisateur du SGCI intéressera particulièrement les concepteurs d'enquête qui ne sont pas certains de la taille de l'ordinateur dont ils auront besoin.

Structures complexes de données

Le SGCI ne peut pas prendre en charge des structures complexes de données, puisqu'il suppose que les données d'enquête à contrôler ont été regroupées sur une seule table ORACLE. Si les données d'enquête sont déjà regroupées en un ensemble de tables ORACLE, on peut utiliser les capacités SQL de réunion et de projection des tables pour créer la table dont le SGCI a besoin. Si les données d'enquête sont extérieures au SGCI, il faudra recourir à un programme personnalisé et créer un "dossier plat" pouvant être chargé dans une table ORACLE.

Données qualitatives et données quantitatives

Le SGCI est surtout axé sur le traitement des données numériques ou quantitatives.

Type d'entrée

Le SGCI accepte un ensemble de règles de contrôle dont le rôle consiste à exprimer les relations existant entre les zones numériques des données d'enquête. Aucune instruction de programmation n'est nécessaire, mais le concepteur d'enquête doit élaborer une stratégie relative à l'ordre dans lequel les données seront contrôlées. Il lui suffit de subdiviser les données d'enquête en ensembles d'enregistrements (strates) et ensuite de traiter les ensembles de variables différents (groupes de contrôle). Il peut également choisir d'appliquer des techniques d'imputation différentes pour corriger des variables différentes. La combinaison de ces différentes techniques de traitement pour arriver à obtenir des données d'enquête exemptes d'erreurs exige une étroite coopération entre les réalisateurs de l'enquête et les responsables des statistiques.

Efficience

Le rendement du SGCI constitue l'inconnue la plus importante. Bien que l'intégration de ce système au SGBD ORACLE comporte de nombreux avantages que nous avons soulignés (notamment vaste dictionnaire de données, transférabilité, etc.), chaque application du système ORACLE entraîne une certaine perte de rendement du SGCI en contrepartie du caractère polyvalent du SGBD. Quant à savoir si cette perte sera trop élevée pour justifier l'emploi du SGCI pour la réalisation d'enquêtes, c'est là une question à laquelle nous ne pouvons répondre.

L'exploitation du SGCI comporte un certain nombre d'étapes exigeant de nombreux calculs. Par exemple, l'algorithme utilisé pour générer des enregistrements fictifs échappant à tous les contrôles mais se trouvant aux extrémités de la région définie comme acceptable par les contrôles linéaires, peut nécessiter la résolution préalable d'un grand nombre d'ensembles d'équations linéaires. L'algorithme qu'utilise actuellement le SGCI est fondé sur les travaux de Chernikova (1965). Si l'utilisateur veut introduire un grand nombre de variables et de règles de contrôle dans un groupe de contrôle, la génération des points extrêmes sera coûteuse.

L'imputation par donneur, fonction non-disponible dans la version du SGCI évaluée par l'auteur, est une opération qui exige également de nombreux calculs. La documentation d'accompagnement indique que l'imputation par donneur est une technique qui exige la construction d'un arbre K-D (Bentley, 1975, 1979), permettant de procéder à la recherche du voisin le plus proche et de trouver ainsi les enregistrements correspondants à l'enregistrement receveur à retenir pour l'imputation. Le temps d'exécution de cette opération dépendra du nombre de zones utilisées pour construire l'arbre et du nombre d'enregistrements donneurs possibles.

Documentation, apprentissage et appui

L'utilisateur qui veut se servir du SGCI pour une enquête doit bien connaître le système ORACLE. En raison de difficultés techniques que n'importe qui peut rencontrer, il a par exemple fallu à l'auteur près de 10 heures pour l'installation initiale du système ORACLE et de SGCI. Le SGCI suppose que les données d'enquête sont déjà regroupées dans une table ORACLE et bien qu'un petit nombre de commandes SQL suffise pour créer une telle table, le novice risque d'être intimidé par l'ampleur de la tâche. Si le système ORACLE peut représenter des problèmes d'apprentissage assez sérieux au départ pour le responsable d'une enquête, l'intérêt d'un tel apprentissage se fera sentir davantage à long terme. Lorsque le personnel affecté aux enquêtes aura appris à se servir d'ORACLE, du SQL et du SGCI, il pourra se servir de ses connaissances pour d'autres enquêtes.

La documentation d'accompagnement du SGCI n'est d'aucune utilité pour le novice. La documentation évaluée par l'auteur comportait 11 documents techniques séparés, dont quatre décrivaient la version 4.0 et étaient par conséquent périmés. L'auteur a également eu accès à plusieurs documents de recherche, portant notamment sur des descriptions d'ensemble du SGCI (Kovar, 1988) et sur la théorie du contrôle numérique (Giles, 1988). Cette documentation ne comportait aucun exemple et aucun échantillon ordinolingue n'était fourni avec le logiciel. Aucune notice explicative n'était prévue à l'intention des novices. Le guide actuel de l'utilisateur explique ce qu'est l'interface avec utilisateur au lieu de dire à quoi il sert et comment s'en servir.

Quand on aura une plus grande expérience de l'exploitation du SGCI pour le traitement des données d'enquête sur la production, il faudra modifier la documentation d'accompagnement afin qu'elle englobe des explications sur la façon dont on peut utiliser ce progiciel pour résoudre différents types de problèmes. La documentation devra également fournir certains résultats sur le rendement du SGCI, afin que l'on puisse déterminer la taille de l'ordinateur nécessaire à telle ou telle application du SGCI.

Coût et valeur

Pour exploiter le SGCI sur un ordinateur personnel IBM, l'utilisateur doit obtenir une licence qui coûte environ 1,500 $ au Canada pour la version 5.1A d'ORACLE. Il doit avoir un ordinateur personnel de marque IBM PC/AT ou compatible de 640K octets de mémoire vive (RAM) et d'au moins 896K octets de mémoire supplémentaire. Le système ORACLE exige environ 7.5M octets de stockage de disque. Le SGCI exige 1M octet d'espace de disque supplémentaire et, de plus, l'utilisateur doit avoir suffisamment d'espace de disque pour que la base de données ORACLE puisse englober à la fois les données d'enquête et les tableaux de commande du SGCI.

À l'heure actuelle, le SGCI n'est pas distribué par Statistique Canada. Il est encore en cours d'élaboration et sa valeur réelle ne pourra être déterminée tant qu'il n'aura pas été exploité pour le traitement de plusieurs enquêtes sur la production.

3.2 Integrated System for Survey Analysis (ISSA) (Système intégré pour l'analyse des enquêtes)

Le ISSA (1988) a été mis au point par le Institute for Resource Development (Institut de développement des ressources) pour son programme d'enquêtes démographiques et sanitaires. Ce programme a été conçu pour appuyer le traitement de 35 enquêtes démographiques et sanitaires de grande envergure, réalisées par l'Institut dans les pays en voie de développement. L'organisation du logiciel permet le traitement en trois étapes des données d'enquête soit entrée, contrôle et tabulation, le tout avec un seul progiciel exploitable sur ordinateur personnel IBM.

La présente évaluation porte sur la version 1.0 du ISSA et s'appuie sur le manuel de l'utilisateur, de 300 pages environ.

3.2.1 Description générale

Lors de l'exécution du programme ISSA, le menu principal est présenté à l'utilisateur; il contient une liste de choix de menus qui apparaît à la dernière ligne au bas de l'écran. Chaque choix de menus s'accompagne d'un numéro de touche de fonction permettant de sélectionner le programme ou l'action souhaitée. Le menu principal comprend les choix suivants:

1. Aide (n'existe pas dans la version 1)
2. Fin du programme ISSA
3. Programme de contrôle du dictionnaire
4. Unité de traitement par lots
5. Programme de production de rapports sur le dictionnaire
6. Programme de gestion de fichiers (n'existe pas dans la version 1)
7. Constitution d'un fichier de transfert pour le SPSS/PC
8. Établissement de listes de renvois au dictionnaire
9. Programme de contrôle des formulaires
10. Entrée des données.

Avant de procéder à une nouvelle application, le concepteur d'enquête doit tout d'abord exécuter le programme de contrôle du dictionnaire ou de contrôle des formulaires pour définir un nouveau dictionnaire de données pour les données de son enquête. C'est pour des raisons historiques qu'il existe deux programmes de contrôle et la documentation d'accompagnement recommande de n'utiliser que le programme de contrôle des formulaires.

Lorsqu'il veut définir un nouveau dictionnaire de données, l'utilisateur voit tout d'abord apparaître à l'écran la description du dictionnaire: la première chose à faire consiste à localiser les variables servant à identifier les différents "niveaux" ou parties de chaque questionnaire. L'utilisateur peut décrire les fichiers plats ou les fichiers hiérarchiques par des "sections" différentes (enregistrements). Dans le cas des fichiers hiérarchiques, l'utilisateur doit également localiser la zone qui contiendra la valeur de l'identificateur de chaque section.

Le programme qui prend en charge les renseignements relatifs au dictionnaire de données est un programme de contrôle qui dépend du mode d'exploitation. Cela permet à l'utilisateur de travailler en mode définition, énumération ou modification. L'utilisateur peut passer d'un mode à l'autre en utilisant une touche de fonction particulière, chaque mode affectant un ensemble différent de significations à chaque touche de fonction. Le menu des significations courantes des touches de fonction apparaît toujours à la dernière ligne, au bas de l'écran.

Une fois que la description du dictionnaire est terminée sur l'écran, elle est suivie par la description des sections, et l'utilisateur peut définir les renseignements suivants pour chaque section :

1. nom de la section et label descriptif
2. code ou identificateur de la section
3. partie ou niveau du dictionnaire auquel appartient la section
4. si la section se retrouve une ou plusieurs fois
5. nombre minimum et maximum de fois où la section peut se retrouver, et
6. nombre minimum et maximum de fois où les variables groupées dans cette section peuvent se retrouver.

Comme l'indiquent les choix sus-mentionnés, le concepteur d'enquête dispose d'une grande souplesse pour décider de la façon dont les sections contiendront les données d'enquête. Par exemple, le système ISSA permet au concepteur d'enquête de constituer un modèle d'un

ensemble de variables, se répétant au moins de deux façons différentes sur un questionnaire. Cet ensemble de variables peut être défini comme faisant partie d'une section multiple ou d'un enregistrement multiple. La section sera alors répétée à chaque apparition de l'ensemble de variables. Les variables peuvent être aussi définies comme constituant ensemble un "groupe", et être placées dans une section qui n'apparaît qu'une fois mais permet la répétition des variables groupées jusqu'à concurrence d'un nombre maximum de fois.

Après avoir défini les renseignements sur les sections, l'affichage à l'écran de la description des variables invite l'utilisateur à fournir des renseignements sur chaque variable de la section. Le système ISSA permet de définir les renseignements suivants pour chaque variable :

1. nom et label descriptif
2. localisation ou longueur de la variable dans la section
3. format (numérique ou alphanumérique)
4. nombre de décimales implicites pour chaque variable numérique
5. si la variable appartient ou n'appartient pas à un groupe de variables
6. valeur spéciale pour les données manquantes
7. valeur spéciale pour les données non-applicables
8. valeur implicite des variables de sortie
9. fourchettes de validité des valeurs
10. labels pour chaque valeur valide des variables.

Lorsque l'utilisateur a défini les renseignements sus-mentionnés pour la variable applicable, il se sert d'une touche de fonction pour mettre à jour le dictionnaire, et ISSA lui présente à nouveau un écran vierge pour la description des variables. Quand toutes les variables d'une section ont été définies, une touche de fonction lui permet de passer à une nouvelle section ou, si toutes les sections sont définies, de sortir du programme de définition du dictionnaire. Le concepteur du système peut alors exécuter le programme de production de rapports sur le dictionnaire, et obtenir un imprimé comportant la liste de tous les renseignements du dictionnaire relatifs aux sections et aux variables.

Pour mettre au point une application ISSA, l'utilisateur doit ajouter d'autres renseignements au dictionnaire de données. Par exemple, pour créer une application entrée de données, l'utilisateur doit d'abord définir un formulaire d'entrée de données pour chaque section de données du questionnaire, par le truchement du programme de contrôle des formulaires. Une fois que le dictionnaire et les formulaires d'entrée de données ont été créés, l'utilisateur doit alors créer les "procédures" ISSA qui vont gouverner l'opération entrée des données. De nouvelles procédures peuvent être ajoutées au dictionnaire de données, grâce au programme de contrôle plein écran du ISSA.

Le langage de programmation du ISSA est un langage hautement structuré. Il n'existe ni programme d'interprétation ni compilateur MS-DOS pour ce langage, et seul le processeur d'entrée des données ou de traitement par lots du ISSA peut définir les énoncés de procédure. La définition des procédures peut se faire pour chaque niveau, section ou variable d'un questionnaire d'enquête. Pour chaque procédure, il suffit d'utiliser les commandes PREPROC (pré-traitement) ou POSTPROC (post-traitement), pour déterminer si la procédure doit être suivie avant ou après que l'on ait trouvé le niveau, la section ou la variable pertinent du questionnaire. Ce type de spécification permet au programmeur de contrôler soigneusement le traitement des fichiers de données hiérarchiques.

Le langage employé pour les procédures ISSA comporte les types de commandes suivants :

1. énoncés d'affectation des valeurs pouvant employer des expressions arithmétiques
2. commande ENCADRÉ pour permettre le recodage des variables

3. commandes de contrôle du programme comme :
   a) IF-THEN-ELSE-ENDIF
   b) PASSER À
   c) WHILE-DO-ENDDO
   d) FIN DE NIVEAU, FIN DE SECTION ET SORTIE.
4. fonctions numériques intégrées comme MOYENNE, COMPTAGE et MAX
5. fonctions intégrées de traitement des fichiers externes.

La liste qui précède est un échantillon des types de commandes ISSA. Ce n'est qu'en étudiant la documentation d'accompagnement du ISSA que l'on peut vraiment apprécier la richesse du langage des procédures ISSA.

Une fois que le dictionnaire des données de base du questionnaire a été modifié pour inclure les formulaires et procédures d'entrée des données, le processeur d'entrée des données permet d'introduire les données d'enquête. Le même processeur peut également servir à vérifier des données introduites antérieurement.

Le concepteur d'enquête peut utiliser le dictionnaire des données de base de l'enquête pour créer d'autres applications ISSA. Par exemple, il peut créer un contrôle secondaire en prenant une copie du dictionnaire de données et en y ajoutant de nouvelles procédures différentes, qui seront exécutées par le programme de traitement par lots du ISSA. Le langage des procédures du ISSA comporte également un ensemble de commandes permettant à l'utilisateur d'exprimer les données d'enquête sous forme de tableaux. On peut créer une application de tabulation en ajoutant de nouvelles procédures employant les commandes de création de tableaux à une autre copie du dictionnaire des données d'enquête, et en soumettant les résultats obtenus au programme de traitement par lots du ISSA.

3.2.2 Évaluation

Fonctions de base et fonctions avancées

Le ISSA fait appel à la majorité des fonctions de base d'un progiciel de contrôle que nous avons énumérées au chapitre 2.1. Le dictionnaire de données ISSA peut être utilisé pour définir le fichier des données d'enquête le plus complexe qui soit, ce qui permet une application entrée de données ISSA garantissant la correction structurelle de chaque questionnaire d'entrée. Les contrôles de validité des variables individuelles peuvent être effectués à partir des spécifications de validité du dictionnaire de données, tandis que le langage de programmation du ISSA servira à vérifier la cohérence intra zone. Le langage de programmation du ISSA ne comporte aucune commande spéciale permettant de modifier automatiquement des données incorrectes, mais il peut prendre en charge les variables et les ensembles à traiter et servir à programmer des solutions personnalisées, comme les "hot decks", par exemple, pour la phase d'imputation des erreurs.

Les fonctions de base d'un progiciel de contrôle que n'offre pas le ISSA sont l'analyse des règles de contrôle et la production de rapports de contrôle. Les données à traiter par le ISSA sont exprimées en langage de programmation et, pour cette raison, l'analyse des règles de contrôle n'est pas possible. Le système ISSA ne produit pas de rapports de contrôle standard mais les types de rapports décrits au chapitre 2.1 peuvent être programmés directement dans le langage de programmation du ISSA.

Le système ISSA offre toutes les fonctions avancées d'un programme de contrôle décrites au chapitre 2.2. L'utilisateur décrit une seule fois les données d'enquête à partir du dictionnaire de données ISSA et combine ensuite une copie de ce dictionnaire de données à différents formulaires et/ou procédures, pour créer les applications d'enquête requises. La nécessité de copier le dictionnaire de données d'enquête constitue un désavantage majeur du ISSA, dans la mesure où chaque changement apporté au dictionnaire de données de base devra être reproduit sur toutes les copies du dictionnaire.

La richesse du langage de définition des données du ISSA entraînera pour certains concepteurs d'enquête un grand nombre de choix difficiles. La première fois que le dictionnaire de données est créé, des décisions doivent être prises quant au nombre des niveaux et des sections à définir par questionnaire et quant à la nécessité de constituer une section multiple ou un groupe de répétitions pour un ensemble donné de variables. Ces décisions sont très importantes car elles conditionnent la forme de l'entrée des données et la façon dont le langage de programmation du ISSA peut servir à la fois à l'entrée des données et au traitement par lots.

Le programme d'entrée des données du ISSA est très puissant et offre toutes les caractéristiques requises, à l'exception des statistiques opérateur. L'une des caractéristiques uniques de son processeur d'entrée de données est sa capacité de garder le contrôle lorsque l'opérateur veut revenir en arrière suite à une instruction "passer à..." de manière à modifier le parcours choisi. Dans ce cas, le ISSA garantit que le code "non-applicable" tel qu'il a été défini par l'utilisateur est affecté aux variables appropriées. La majorité des autres progiciels d'entrée de données permettent à l'opérateur de modifier n'importe quelle variable unique sur l'écran, mais ne se préoccupent pas de l'impact que cela peut avoir sur les autres variables.

Le langage de programmation du ISSA peut être utilisé pour créer des applications de tabulation et des fichiers de données extraites à partir du fichier de données d'origine.

Intégration

Le principal objectif de conception du ISSA était d'intégrer les trois principales opérations de traitement d'une enquête. Cet objectif a été atteint avec succès. Le ISSA présente un interface utilisateur cohérent pour chacun des modules du programme, la majorité des entrées étant assurées par le remplissage des zones et la majorité des actions étant spécifiées par le choix de la touche de fonction appropriée, à partir d'un menu affiché à l'écran. Lorsqu'une entrée à format libre est nécessaire pour définir les formulaires ou procédures ISSA, le ISSA fait appel à un programme de contrôle spécialisé plein écran qui se charge de l'opération. Les utilisateurs peuvent également créer leurs propres procédures ISSA avec leur programme de contrôle DOS favori, et les faire passer ensuite directement sous le contrôle du ISSA.

En raison des nombreuses capacités différentes du ISSA, les novices auront peut- être de la difficulté à mémoriser le rôle des différentes touches de fonction pour chaque module du programme. La version 1 du ISSA n'offre pas de fonction d'aide en direct susceptible de résoudre ce problème.

Stratégie de contrôle et mode d'exploitation

Le ISSA permet au concepteur d'enquête de choisir à quel moment les données d'enquête doivent être corrigées. En ajoutant des procédures ISSA sophistiquées à l'application entrée des données d'enquête, l'ensemble du contrôle des données d'enquête peut se faire au moment de l'entrée initiale. Si le questionnaire d'enquête est très complexe et si les opérateurs chargés de l'entrée des données ont l'habitude de s'en servir, c'est sans doute la meilleure technique de correction des erreurs. D'autre part, le concepteur d'enquête peut également utiliser le ISSA pour créer un programme d'entrée de données qui ne fera que contrôler la validité de zones individuelles. Le contrôle des données d'enquête résultantes ne pourra être effectué qu'en créant une application ISSA séparée de traitement par lots.

Transférabilité

Le ISSA est rédigé en langage de programmation C (Kernighan, 1978); il est exploitable sur ordinateur personnel IBM et n'est pas à l'heure actuelle utilisable sur un autre ordinateur ou avec un autre système d'exploitation.

Structures complexes de données

Le dictionnaire de données du ISSA peut permettre de décrire jusqu'au plus complexe des questionnaires d'enquête. Le concepteur d'enquête a la possibilité de diviser un questionnaire en différents niveaux et sections (enregistrements) et de définir des groupes de répétition des variables. Le langage de programmation du ISSA permet également au programmeur de se référer aux fichiers externes qu'il aura pu constituer à partir d'une application ISSA séparée. Ainsi, le concepteur d'enquête peut répartir les données d'une enquête en plusieurs fichiers séparés, tout en continuant à les traiter ensemble.

Données qualitatives et données quantitatives

Au départ, le ISSA a été conçu pour traiter les données d'enquêtes sanitaires et démographiques de grande envergure, pour lesquelles la majorité des zones sont qualitatives. Le dictionnaire de données du ISSA permet de définir de multiples fourchettes de validité pour les variables, tandis que le langage de programmation comporte plusieurs commandes spécifiquement axées sur le traitement des variables qualitatives. Aucune caractéristique spéciale n'est prévue pour le traitement des données quantitatives.

Type d'entrée

La conception d'un dictionnaire de données ISSA et ses applications connexes doivent être confiées à un spécialiste du traitement des données. Bien que le programme de contrôle des formulaires soit en soi relativement simple à exécuter, les décisions à prendre pour mettre au point le traitement d'un questionnaire d'enquête complexe exigent habituellement les compétences d'un analyste de systèmes. En outre, le langage de programmation du ISSA est plutôt complexe et seul un programmeur pourra l'employer avec succès.

Efficience

Bien que le langage des commandes du ISSA soit interprété par le programme d'entrée des données ou le programme de traitement par lots, les applications que l'auteur cite en exemple sont exécutées extrêmement rapidement. Si l'opérateur recherche un fichier de données créé par le ISSA, il trouvera très rapidement un questionnaire particulier, chaque dossier étant automatiquement indexé. Dans le cas des grands fichiers de données, les opérations de tabulation peuvent être lentes, quoiqu'il semble que le système soit largement tributaire de la vitesse d'exécution des périphériques d'entrée/sortie. Seule une structure sous-jacente de données différente, comme les fichiers transposés (Cotton, 1979), est susceptible d'accélérer l'exécution des applications de tabulation.

Documentation, apprentissage et appui

Le manuel de documentation de la version 1.0 du ISSA a plus de 300 pages. Il fait à la fois office de guide d'introduction pour l'utilisateur et de manuel de référence du ISSA. C'est un document de référence adéquat et utile pour un expert, puisqu'il contient une définition de toutes les capacités du ISSA. Il comprend une table des matières principale et chacun de ses 15 chapitres est précédé d'un index détaillé. L'évaluateur aurait préféré qu'il n'y ait qu'un seul index général et plus complet, permettant de se renseigner sur la signification de telle ou telle touche de fonction ou la syntaxe exacte d'une commande du ISSA. Le manuel devrait comporter un plus grand nombre d'exemples d'utilisation du langage des commandes du ISSA.

Malheureusement, il ne présente aucune utilité pour l'utilisateur novice. Chaque chapitre décrit en détail une certaine partie du ISSA, ce qui risque de dépasser complètement un novice. Pour aggraver encore la difficulté, de nombreuses parties du manuel renvoient à un questionnaire complexe de 13 pages relatif à une enquête sur la santé, et inclus en annexe.

Il faudrait que l'utilisateur novice du ISSA puisse se familiariser par étapes avec le programme, grâce à des exemples beaucoup plus simples et ne faisant pas immédiatement appel à toute la richesse du progiciel ISSA. Il serait extrêmement utile de mettre au point séparément un guide de l'utilisateur, s'appuyant sur un questionnaire plus simple et des exemples directs.

Le système ISSA n'est pas à l'heure actuelle doté d'une capacité d'aide en direct, mais d'après la documentation traitant de l'ensemble des programmes, ce service sera assuré par la touche de fonction F1. Lorsque cette capacité sera disponible, les utilisateurs novices la trouveront indéniablement très utile.

Pour les nouveaux utilisateurs, il serait utile que la version ordinolingue des exemples cités dans la documentation d'accompagnement du ISSA soit fournie. Les exemples devraient être présentés au complet et inclure non seulement un exemple d'application d'entrée de données avec le dictionnaire de données de base correspondant, mais encore un échantillon des fichiers de données et des descriptions des procédures et formulaires utilisés pour plusieurs applications ISSA.

Coût et valeur

Le logiciel et la documentation d'accompagnement du ISSA sont disponibles au Canada au coût d'environ 500$. Il n'est pas nécessaire d'acquérir un autre logiciel pour exploiter le ISSA sur un ordinateur personnel IBM. Le progiciel ISSA représente un rapport qualité/prix très intéressant, dans la mesure où on peut l'utiliser à toutes les étapes du traitement des données d'enquête.

3.3 PCEDIT

PCEDIT (1988) est un progiciel "intelligent" d'entrée et de contrôle des données, exploitable sur ordinateur personnel IBM. Il est disponible au Département de la coopération technique pour le développement des Nations-Unies. Ce progiciel a été élaboré dans le cadre du projet Logiciel et soutien pour le traitement des données démographiques, projet qui a déjà permis la distribution des progiciels statistiques UNEDIT, XTALLY et PC-COXTALLY. Ces progiciels ont tous été élaborés de façon à apporter des solutions peu coûteuses aux problèmes du traitement des données démographiques, en particulier pour les utilisateurs et organismes de statistique oeuvrant dans les pays les moins avancés.

La présente évaluation porte sur la version 2.11 du PCEDIT et sur la documentation d'accompagnement de la version 2.10, la version mise à jour de la documentation d'accompagnement de la version 2.11 n'étant pas disponible.

3.3.1 Description générale

Selon la documentation d'accompagnement du PCEDIT, ce programme permet l'entrée "intelligente" des données. L'utilisateur a la possibilité de décrire la définition des fichiers relatifs au questionnaire ainsi que l'ensemble des règles de contrôle applicables, ce qui permet au PCEDIT de commander l'entrée, le contrôle et la modification des données du questionnaire. Le PCEDIT recueille également des statistiques au moment de l'entrée des données et de la vérification de chaque lot de questionnaires.

Toutes les capacités du PCEDIT sont activées par dialogue interactif avec l'utilisateur. Lorsque le PCEDIT est appelé pour la première fois, le menu principal apparaît et l'utilisateur peut choisir parmi les sept fonctions suivantes :

1. Mise à jour de la définition des fichiers
2. Programme d'entrée des données
3. Programme de vérification

4. Programme de modification
5. Programme de remplissage de zéros
6. Renseignements sur les fichiers de données
7. Paramètres du système

La fonction définition des fichiers permet à l'utilisateur de décrire un nouveau questionnaire ou de modifier une définition existante. Un questionnaire PCEDIT peut comporter jusqu'à 64 types d'enregistrement différents. Bien qu'il n'y ait pas de limite au nombre de zones d'un enregistrement, en pratique ce nombre est limité du fait qu'un enregistrement ne peut pas dépasser 80 caractères.

Pour la description d'un questionnaire à enregistrements multiples, la première zone de chaque enregistrement est celle qui détermine le type d'enregistrement. Une fois que la zone est décrite, le PCEDIT en affiche la définition en indiquant, en haut de l'écran, le type d'enregistrement, la position de départ de la zone en cours de traitement et la longueur totale de l'enregistrement. La dernière ligne au bas de l'écran fait le point de la situation et comporte un court message d'aide, indiquant à l'utilisateur ce qu'il doit faire ensuite. Cette ligne est une constante efficace de toutes les fonctions du PCEDIT.

Pour chaque zone en cours de traitement, le PCEDIT demande à l'utilisateur de lui donner le nom de la zone, le type (alphanumérique ou numérique), la longueur et la situation de la vérification (OUI ou NON). Le PCEDIT suppose qu'une nouvelle zone commence immédiatement là où se termine la précédente, si bien que la position de départ d'une zone est déterminée automatiquement. Bien que l'utilisateur soit obligé de définir les zones d'un enregistrement en ordre séquentiel, les erreurs peuvent être corrigées avec certains efforts, puisque l'utilisateur peut supprimer des zones et en ajouter de nouvelles.

Après avoir introduit les descriptions des zones, l'utilisateur peut ensuite introduire jusqu'à 20 règles de contrôle pour la zone en cours de traitement. Ces règles sont basées sur un langage simple à expression linéaire et à mots clés. Dès que l'utilisateur est prêt à introduire une nouvelle règle de contrôle, la ligne qui fait le point de la situation indique quels sont tous les mots clés valides. Il suffit de taper le premier caractère d'un mot clé pour que le reste du mot apparaisse automatiquement. Le PCEDIT demande alors à l'utilisateur de lui donner le mot clé valide suivant, le nom de l'opérateur ou le nom de la zone. C'est la structure rigide du langage d'entrée qui permet l'élaboration interactive des règles.

Les règles de contrôle applicables à une zone peuvent spécifier une vérification de contrôle de la zone, ce qui permet de déterminer si les données introduites se trouvent à l'intérieur de la fourchette valide des valeurs, et(ou) un contrôle de la cohérence de la zone par rapport aux autres zones déjà introduites de l'enregistrement. S'il rencontre des valeurs particulières, l'utilisateur peut spécifier que le programme d'entrée des données demande une nouvelle valeur valide, de passer à une autre zone, de passer à l'enregistrement suivant ou encore d'imprimer un message relatif aux erreurs. À l'heure actuelle, l'utilisateur ne peut pas spécifier de règles de contrôle structurel.

3.3.2 Évaluation

Fonctions de base et fonctions avancées

Le PCEDIT n'est pas un progiciel de contrôle complet, dans la mesure où il n'offre pas toutes les fonctions de base d'un programme de contrôle (voir chapitre 2.1). En réalité, le PCEDIT n'est rien d'autre qu'un progiciel d'entrée des données qui permet au concepteur d'enquête de décrire le fichier des données d'enquête et de programmer les contrôles de la validité des valeurs et les contrôles de cohérence. À l'heure actuelle, il ne permet ni de procéder à des contrôles structurels, ni d'analyser les règles de contrôle, et le seul fichier de sortie qu'il

peut générer résulte de l'opération entrée de données. Il ne comporte aucune capacité de correction automatique des erreurs, c'est-à- dire qu'il n'a aucune capacité intégrée de traitement "hot deck" en langage de programmation. Les seuls rapports de contrôle qui sont générés signalent le nombre d'erreurs de frappe commises par l'opérateur et le nombre de corrections requises.

Le PCEDIT offre la majorité des caractéristiques principales d'un progiciel d'entrée des données d'enquête, que nous avons décrites au chapitre 2.2. Son plus grand inconvénient est que l'utilisateur n'a pas la possibilité de créer des écrans d'entrée des données. L'entrée des données est linéaire et le schéma de frappe d'entrée des données de l'opérateur qui apparaît à l'écran est exactement le même pour chaque enregistrement. Le PCEDIT peut prendre en charge la capacité de détecter les erreurs d'entrée, de programmer un schéma d'instructions de "passer à...", de répéter automatiquement les zones ou de programmer la mise en séquence des zones introduites. Le PCEDIT est doté d'une capacité de vérification des données et permet à l'utilisateur d'extraire et de modifier un questionnaire.

La capacité de dictionnaire des données du PCEDIT permet à l'utilisateur de nommer les zones, de décrire leur type et de spécifier si elles doivent ou non être vérifiées. Le PCEDIT n'a pas la capacité de nommer des valeurs codées.

Intégration

Le PCEDIT offre un interface utilisateur cohérent, et son système de menus hautement structuré de même que son système de guidage de l'opérateur en fonction du contexte en font un programme dont il est facile d'apprendre à se servir. Quand les auteurs intégreront de nouvelles fonctions au PCEDIT, ils devront s'efforcer de préserver sa cohérence.

Stratégie de contrôle et mode d'exploitation

Le concepteur qui opte pour le PCEDIT choisit un progiciel autorisant uniquement le contrôle interactif des données d'enquête au moment de l'entrée ou de la vérification des données. Il n'a aucune capacité de contrôle par lots.

Transférabilité

La conception du PCEDIT est surtout axée sur la vitesse d'exécution. À cette fin, le progiciel a été codé dans le langage d'assemblage du IBM PC. Il n'est donc pas transférable, sauf peut-être sur un autre système d'exploitation IBM PC, comme le OS/2, par exemple.

Structures complexes de données

Comme nous l'avons vu au chapitre 3.3.1, le PCEDIT peut traiter un questionnaire comportant jusqu'à 64 types d'enregistrement différents. Il n'a aucune capacité de définition ou d'utilisation de groupes de variables répétés.

Données qualitatives et données quantitatives

Ce progiciel n'offre aucune capacité spéciale de traitement des données quantitatives. Le langage de programmation des contrôles rend extrêmement simple le contrôle des multiples fourchettes de validité des valeurs d'une variable qualitative.

Type d'entrée

Le PCEDIT a été rédigé de façon à ce qu'un non-programmeur puisse effectuer la définition du questionnaire et spécifier les règles de contrôle. D'où sa grande simplicité d'exploitation et l'existence à toutes les étapes d'un système de guidage de l'utilisateur selon le contexte. Il faudra attendre que le PCEDIT ait été utilisé pour traiter des données réelles d'enquête, pour déterminer dans quelle mesure les utilisateurs finals pourront s'en servir avec succès.

Efficience

L'exécution est instantanée pour pratiquement toutes les fonctions du PCEDIT sauf s'il s'agit de rechercher un fichier de données dans le but de retrouver un questionnaire spécifique. Cette fonction peut prendre du temps, ce qui rend extrêmement long le processus de modification de certains questionnaires spécifiques. Comme pour le ISSA, il pourrait être utile de prévoir pour le PCEDIT la constitution d'un fichier indexé permettant d'accéder rapidement et directement à tel ou tel questionnaire spécifique.

La majorité des opérateurs d'entrée de données s'attendront à pouvoir utiliser le clavier numérique du IBM PC. Malheureusement, le PCEDIT se sert des touches marquées d'une flèche vers la gauche et vers la droite pour passer d'une zone à l'autre. Il faudra peut-être modifier cela de façon à ce que les touches TAB et SHIFT TAB puissent servir à passer d'une zone à l'autre, le clavier numérique pouvant alors être exclusivement réservé à l'entrée des données.

Documentation, apprentissage et appui

Le logiciel et la documentation d'accompagnement du PCEDIT sont actuellement disponibles en français et en anglais. Le progiciel comprend également un programme de démonstration qui présente à l'utilisateur un échantillon des fonctions définition des fichiers, entrée des données et vérification.

La documentation d'accompagnement écrite du PCEDIT est très faible et elle exige énormément de travail de la part de l'utilisateur. La version anglaise n'a que 23 pages et si elle souffre de nombreuses erreurs de grammaire et d'orthographe, elle ne comporte aucun exemple et aucun échantillon des images-écran. L'explication de la syntaxe des règles de contrôle est très compliquée à comprendre, dans la mesure où les règles sont décrites par la première lettre du mot clé plutôt que par le mot clé tout entier. Le manuel et le programme de démonstration devraient être coordonnés et reprendre le même exemple simple. Il serait bon de présenter la définition des fichiers de même que certains fichiers réels de données, pour étayer ce type d'application. Si possible, il serait également souhaitable de prévoir un exemple plus complexe et différent, à l'intention des experts ou des professionnels du traitement des données.

La documentation écrite devrait également contenir une description générale du pourquoi et comment utiliser le PCEDIT. Elle devrait préciser les étapes à franchir ainsi que les décisions stratégiques à prendre à chaque étape.

Le programme de démonstration du PCEDIT est assez bon mais il est trop court et ne contient pas suffisamment d'exemples de règles de contrôle. Le premier écran du programme de démonstration indique comment avancer et revenir en arrière dans la démonstration, mais ne mentionne pas s'il est possible de reprendre la démonstration à un point précis autre que le commencement.

Les capacités d'aide en direct du PCEDIT constituent l'un de ses principaux atouts. L'utilisateur peut toujours déterminer les fonctions offertes en se référant à la ligne qui fait le point de la situation ou en tapant sur une touche de fonction faisant apparaître à l'écran une case dans laquelle est indiquée la liste des actions pilotées par chaque touche de fonction, ou encore les séquences d'action pilotées par la touche CONTROL.

Bien que l'auteur du progiciel PCEDIT se soit montré très coopératif pendant l'évaluation, il n'existe à l'heure actuelle aucune méthode d'appui officielle.

Coût et valeur

Le PCEDIT est distribué gratuitement par les Nations-Unies. Cela en fait un progiciel d'entrée des données peu coûteux pour une application d'enquête ou de recensement exigeant

l'utilisation de plusieurs IBM PC. Le PCEDIT peut facilement tolérer les taux élevés d'entrée des données que l'on peut attendre d'un recensement, soit plus de 10,000 frappes à l'heure.

On peut également envisager d'exploiter le PCEDIT pour l'entrée et le contrôle en direct des données d'un questionnaire d'enquête complexe. Mais si le concepteur d'enquête recherche des contrôles le moindrement complexes, il devra recourir à un progiciel de contrôle plus complet.

3.4 Integrated Microcomputer Processing System (IMPS) (Système intégré de traitement sur micro-ordinateur)

Le IMPS a été élaboré par le International Statistical Programs Center of the U.S. Census Bureau (Centre international des programmes statistiques du Bureau américain du recensement). Il s'appuie sur l'expérience acquise par le Centre en matière de traitement des données de recensement sur micro-ordinateurs, dans les pays en voie de développement. Le système IMPS intégral que décrivent Toro et Chamberlain (1988) comprendra les principales composantes suivantes: un dictionnaire de données (DATADICT, 1987), un module d'entrée de données (CENTRY), un progiciel de contrôle (CONCOR 1988), un progiciel de tabulation (CENTS 1988), un système de gestion et de rapports sur le recensement (dCONTROL) et un progiciel de planification du recensement (CENPLAN). La présente évaluation ne porte que sur les trois premiers de ces modules, lesquels servent à l'entrée et au contrôle des données de recensement de l'utilisateur.

Parmi les logiciels exploitables sur ordinateur personnel IBM, nous nous sommes penchés sur les suivants : a) DATADICT version 1.0, b) progiciel de démonstration de CENTRY et c) CONCOR version 3.0

3.4.1 Description générale

Le IMPS a été élaboré au cours des trois dernières années grâce aux efforts soutenus du Centre international des programmes statistiques (ISPC) en matière de mise à jour et d'amélioration des logiciels que le IMPS peut recevoir. En 1985, le Centre a transféré les logiciels CONCOR et CENTS sur ordinateur personnel IBM. À cette époque, ces deux logiciels exigeaient deux étapes séparées de définition des données, même si les extrants de CONCOR étaient fréquemment utilisés directement par CENTS. À peu près à la même époque, le Centre a effectué une évaluation des progiciels commerciaux d'entrée de données exploitables sur ordinateur personnel IBM et susceptibles de servir à l'entrée des données d'enquête ou de recensement. Naturellement, pour chacun de ces logiciels, le concepteur d'enquête doit effectuer une troisième définition indépendante et différente des données.

L'apparition d'un dictionnaire de données commun tel que celui qui est défini dans le programme DATADICT a permis l'intégration de CONCOR au programme d'entrée de données CENTRY, récemment mis au point. DATADICT peut être utilisé sur ordinateur personnel IBM soit en mode interactif, soit en mode de traitement par lots. En mode interactif, la machine demande les renseignements nécessaires à l'utilisateur et produit un fichier des énoncés des lots, lequel est ensuite traité pour donner un fichier du dictionnaire de données pouvant être lu directement par CENTRY et CONCOR. S'il le désire, l'utilisateur peut créer directement les énoncés des lots DATADICT en utilisant un programme de contrôle de texte.

DATADICT permet à l'utilisateur de décrire la définition des données d'enquête ou de recensement à traiter. En mode interactif, l'utilisateur est guidé vers ce qu'il doit faire. Le formulaire pleine page qui apparaît à l'écran lui permet de définir les renseignements suivants sur le niveau de l'enquête :

1. nom du dictionnaire
2. longueur de l'enregistrement le plus long
3. valeurs implicites non rapportées et non applicables s'appliquant à toutes les variables.

La dernière ligne au bas de l'écran indique les touches de fonction ou les touches spéciales qui sont actives. Par exemple, en frappant sur la touche de fonction 1, l'utilisateur peut obtenir en direct des indications sur ce que signifie l'image-écran qu'il a sous les yeux. Il peut obtenir directement de l'aide sur n'importe quelle zone en frappant la touche de fonction 2 après avoir positionné le curseur sur la zone appropriée. Le menu indique également à l'utilisateur comment passer à l'écran suivant ou revenir à l'écran précédent.

Les images-écran suivantes demandent à l'utilisateur de décrire la position et le type de la zone appartenant à tel ou tel type d'enregistrement (numérique ou alphanumérique) et de donner un nom et une valeur à chaque type d'enregistrement différent. L'utilisateur doit ensuite définir les variables "communes" à chaque enregistrement. Il s'agit habituellement de variables géographiques ou de variables d'identification d'un questionnaire. Finalement, l'utilisateur doit définir les variables traitées pour chaque type d'enregistrement different. Les renseignements requis comprennent, entre autres :

1. le nom complet de la variable, à partir duquel un nom abrégé est automatiquement dérivé
2. la longueur et la position de départ de la variable,
3. si la variable est ou n'est pas une subdivision d'une variable plus grande
4. le nombre de fois où la variable répétée apparaît
5. le type de la variable (numérique ou alphanumérique), et
6. les valeurs de validité et les labels correspondant aux valeurs de la variable.

Quand l'étape interactive d'entrée des données est terminée, DATADICT génère automatiquement le fichier de dictionnaire des données, qui peut être utilisé avec CONCOR ou CENTRY.

Le système CENTRY donne au concepteur d'enquête la possibilité de développer une application entrée des données permettant l'entrée, la vérification et la modification des données de recensement, ainsi que la collecte de statistiques opérateur. Le programme CENTRY comprend un module qui permet au programmeur de définir l'application entrée de données, et un autre module qui permet aux opérateurs de l'entrée des données d'exécuter l'application.

Le module CENTRY destiné au concepteur d'enquête utilise comme entrée la définition des données du fichier de données d'enquête produit par DATADICT. Il permet à l'utilisateur de "reproduire" à l'écran un formulaire d'entrée de données grâce au programme de contrôle plein écran. L'utilisateur peut ainsi concevoir différents formulaires pour les renseignements traités par lots, pour les renseignements sur le niveau du questionnaire et pour chaque type d'enregistrement différent du questionnaire. L'utilisateur peut facilement contrôler la séquence d'entrée des zones et les dupliquer automatiquement. Le concepteur de l'application n'est pas tenu de définir les valeurs de validité de chaque variable, dans la mesure où elles sont déjà définies dans le dictionnaire de données. Le programme CENTRY ne permet pas d'effectuer des contrôles de cohérence d'une zone à l'autre, ces types de contrôle n'étant pas souhaitables dans le cadre d'un recensement, en raison du caractère primordial de la rapidité d'exécution de l'entrée des données.

Le module d'exécution CENTRY prend en charge la description de l'application créée par le module destiné au concepteur d'enquête et permet à l'opérateur d'entrée des données de créer, de modifier et de vérifier les lots des données d'enquête. Le module d'exécution CENTRY fonctionne très rapidement et peut être utilisé sur ordinateur personnel IBM à 256K octets de mémoire et deux unités de disques souples.

Le IMPS dépend du système CONCOR pour la détection et la correction des erreurs des données d'enquête ou de recensement. Le système CONCOR a été utilisé dans plus de 100 pays pour le traitement de données sur la population et le logement émanant de recensements nationaux, de même que pour le traitement de données de recensement sur l'agriculture, d'études sur la population active et d'enquêtes sur l'éducation. Le programme CONCOR est rédigé en langage COBOL et il fonctionne non seulement sur ordinateur personnel IBM mais encore sur un grand nombre d'ordinateurs centraux et de mini-ordinateurs.

La version 3.0 du CONCOR, exploitable sur ordinateur personnel IBM seulement, utilise le fichier du dictionnaire de données créé par DATADICT pour décrire le fichier d'entrée à traiter. Le dictionnaire de données se combine avec les énoncés CONCOR de l'utilisateur, et s'il n'y a pas d'erreur de syntaxe, il est possible de générer un programme source rédigé en COBOL, et ensuite de le compiler sur le compilateur COBOL approprié. Le programme de contrôle de l'utilisateur est capable de lire un fichier de données d'enquête, de modifier les données conformément aux instructions de contrôle CONCOR de l'utilisateur et de produire un fichier de sortie de données exactes de même que n'importe quelle statistique de contrôle demandée.

Le langage de programmation du CONCOR est un langage hautement structuré, spécialement conçu pour la rédaction d'instructions de contrôle informatique. Il comprend des énoncés de structure et des énoncés d'exécution. Les énoncés de structure permettent au programmeur de contrôler l'exécution d'ensemble du programme de contrôle de l'utilisateur, et à ce dernier de définir les éléments suivants :

1. le dictionnaire de données à utiliser
2. la région géographique pour laquelle des rapports de contrôle doivent être créés
3. l'ensemble des variables qui identifient de façon unique un questionnaire
4. le nombre maximum de chaque type d'enregistrement que l'on s'attend à trouver dans un questionnaire
5. que les rapports de contrôle standard doivent être générés
6. les types de contrôles spéciaux à exécuter, par exemple division par zéro
7. les variables à traiter qui seront utilisées dans les énoncés d'exécution CONCOR.

Le langage CONCOR comporte un grand nombre d'énoncés d'exécution différents, mais on peut grosso modo les diviser selon les catégories suivantes :

1. test de condition
2. affectation ou transformation des données
3. imputation des données
4. organisation et commandes du programme, et
5. création d'extrants.

Chaque programme CONCOR se divise en trois parties essentielles. Le sous-programme PROLOG est exécuté une seule fois au début du programme, le sous-programme EPILOG est exécuté une seule fois à la fin du programme et le sous-programme FILTER est exécuté pour chaque type d'enregistrement d'un questionnaire. Les sous-programmes FILTER sont des circuits "automatiques" et leurs énoncés d'exécution sont répétés une seule fois pour chaque type d'enregistrement différent d'un questionnaire.

3.4.2 Évaluation

Fonctions de base et fonctions avancées

Le IMPS offre la majorité des fonctions de base d'un programme de contrôle décrites au chapitre 2.1. Le module DATADICT permet à l'utilisateur de décrire la définition du fichier

de données d'enquête tandis que CONCOR se charge de la majorité des autres fonctions de base, et notamment : contrôles de validité, contrôles de cohérence, correction automatique, production d'un fichier de données corrigées et production de rapports de contrôle. Le système CONCOR ne peut pas analyser les règles de contrôle, puisqu'il fait appel à un langage de programmation et non à un ensemble de règles de contrôle à vérifier. CONCOR ne peut pas produire de fichiers de données extraites puisque le fichier de sortie doit avoir le même format que le fichier d'entrée.

Les modules DATADICT et CENTRY du programme IMPS exécutent deux fonctions avancées d'un programme de contrôle. Ces deux modules présentent presque toutes les caractéristiques requises, à l'exception de celles que nous venons de citer. Bien que DATADICT permette à l'utilisateur de définir des labels pour les codes des variables qualitatives, le programme CONCOR de l'utilisateur ne peut pas ensuite renvoyer à ces labels. CENTRY ne permet pas de réaliser des contrôles de cohérence intra-zone, dans la mesure où il sert surtout pour des applications de recensement, pour lesquelles la rapidité de l'entrée des données est essentielle. CENTRY ne permet pas non plus de donner l'instruction de "passer à..." ni d'insérer automatiquement des codes non applicables pour les variables que l'on a passées.

D'après notre évaluation, le IMPS n'a ni capacité de tabulation ni d'extraction. Cependant, pour un recensement, le concepteur d'enquête peut utiliser le système IMPSCENTS pour produire des tabulations instantanées.

Intégration

Force est de reconnaître que les composantes du IMPS ne sont pas hautement intégrées, puisqu'elles utilisent des interfaces utilisateur et des langages d'entrée différents. En mode interactif, ce sont les mêmes touches de fonction qui permettent d'obtenir de l'aide en direct pour chaque module, mais le menu en cours n'apparaît pas au même endroit. Cela ne constitue pas un désavantage majeur, mais des interfaces plus cohérentes faciliteraient l'apprentissage des novices.

En outre, l'intégration du IMPS serait bien meilleure s'il n'y avait qu'un seul système de menu principal, permettant à l'utilisateur de passer sans difficulté d'un module à l'autre.

Stratégie de contrôle et mode d'exploitation

La stratégie de contrôle du IMPS part de l'hypothèse que ce programme servira uniquement à des enquêtes de très grande envergure et exigeant de très nombreuses entrées de données. Les contrôles de validité ou des fourchettes de validité effectués par CENTRY ne portent que sur des variables uniques - toutes les autres activités de contrôle exigeant l'exécution par lots des données d'enquête par l'entremise du programme de contrôle CONCOR. Alors que les programmes DATADICT et CENTRY peuvent fonctionner en mode interactif, CONCOR est intrinsèquement un processus de traitement par lots. L'utilisateur crée un programme CONCOR au moyen d'un programme de contrôle de texte et soumet ensuite les résultats au système CONCOR pour analyse. Lorsque l'utilisateur dispose d'un programme dénué de toute erreur de syntaxe, le système CONCOR peut générer un programme source rédigé en COBOL, qui doit à son tour être compilé et connecté, pour créer le programme d'exécution de l'utilisateur s'il s'agit d'un programme CONCOR complexe. Chacune de ces étapes peut prendre plusieurs minutes. L'inconvénient de cette méthode de compilation-exécution est compensé par le fait que le programme d'exécution qui en résulte est très rapide.

Transférabilité

Les progiciels DATADICT et CENTRY fonctionnent uniquement sur ordinateur personnel IBM. Bien que la version 3.0 du CONCOR ne soit pas exploitable sur les plus gros modèles, la version 2.4 qui est étroitement compatible avec la version 3.0, peut fonctionner sur de

nombreux ordinateurs centraux et mini-ordinateurs. Il est très possible de concevoir une application IMPS sur ordinateur personnel IBM, et de se servir de CONCOR 2.4 sur un ordinateur plus important pour exécuter le contrôle de la production.

Structures complexes de données

DATADICT permet au concepteur d'enquête de décrire un fichier plat ou hiérarchique contenant jusqu'à 50 types d'enregistrement différents. La seule restriction est que pour chaque enregistrement, l'indication du type doit se trouver au même endroit, et chaque enregistrement doit avoir une taille fixe. DATADICT permet également la répétition de groupes de variables ou de variables individuelles, pouvant ensuite être traités comme des ensembles de variables par les modules du IMPS.

Données qualitatives et données quantitatives

La plupart des énoncés d'exécution pris en charge par CONCOR sont surtout axés sur le contrôle des données qualitatives. Il n'existe aucune capacité spéciale de contrôle des données quantitatives.

Type d'entrée

Les progiciels DATADICT et CENTRY sont assez faciles à utiliser. Un expert ayant des notions minimum en matière de traitement des données mais connaissant bien le questionnaire de l'enquête ou du recensement devrait être en mesure de créer un dictionnaire de données IMPS et de réaliser une application entrée de données CENTRY en moins de deux jours de travail. Quant à l'élaboration des instructions de contrôle CONCOR, elle exige une solide expérience de la programmation. Il faut cependant souligner que le plan et la stratégie globale de contrôle sont plus importants que les énoncés CONCOR. Aucune programmation CONCOR, aussi sophistiquée soit-elle, ne pourra remplacer une stratégie de contrôle globale et bien pensée.

Efficience

Les applications entrée de données du CENTRY sont très rapides et pourront tolérer sans difficulté des taux d'entrée très élevés. Une fois que le programme CONCOR de contrôle de l'enquête a été rédigé et testé, la vitesse d'exécution est très rapide. L'auteur a évalué une application échantillon de recensement : il a pu traiter plus de 10 000 enregistrements minute sur un ordinateur de 20 Mhz compatible avec l'ordinateur IBM PC.

Documentation, apprentissage et appui

La documentation d'accompagnement du IMPS varie selon les composantes. La documentation d'accompagnement du CONCOR est excellente et comprend un guide de l'utilisateur, un manuel de mise au point et un manuel de référence. Le guide de l'utilisateur répond directement aux besoins d'un novice; il présente un exemple complet et plusieurs fragments de codes différents permettant de mettre en lumière les diverses possibilités offertes au programmeur. Un exemplaire ordinolingue de l'exemple au complet est fourni avec le progiciel CONCOR. Le manuel de l'utilisateur comprend également plusieurs chapitres d'introduction très utiles et bien écrits sur les exigences du contrôle en général, ainsi qu'un chapitre final qui récapitule les étapes nécessaires à l'élaboration d'un programme CONCOR satisfaisant.

La documentation DATADICT se présente sous la forme d'un manuel de 52 pages qui fait à la fois office de guide de l'utilisateur et de manuel de référence. Bien que le programme interactif DATADICT offre une capacité d'aide en direct, il serait souhaitable d'améliorer le manuel écrit, qui ne comprend à l'heure actuelle aucune présentation sous forme d'écran,

afin d'inclure un échantillon des séquences ainsi que les reproductions appropriées des images-écran. Au moment où nous avons effectué la présente évaluation, la documentation écrite CENTRY n'était pas disponible.

Coût et valeur

On peut se procurer les modules du IMPS auprès du International Statistical Programs Center (Centre international des programmes statistiques), à un coût nominal. Les modules CONCOR (et CENTS) du IMPS exigent l'obtention d'une licence pour utiliser le compilateur REALIA COBOL (1988) sur IBM PC, ce qui coûte environ 1,500$ au Canada.

4. Conclusions

Les progiciels de contrôle des données d'enquête que nous avons évalués dans le présent document témoignent de l'impact de la haute technologie sur le traitement des données d'enquête et de recensement. Chacun de ces progiciels permet au concepteur d'enquête d'exécuter sur ordinateur personnel IBM certaines ou l'ensemble des fonctions de traitement d'une enquête ou d'un recensement. La capacité d'utiliser le mode interactif et personnel d'un IBM PC pour le traitement des données d'enquête représente un progrès majeur qui n'a été réalisé qu'au cours des deux ou trois dernières années.

Plusieurs facteurs revêtent une importance capitale dans l'apparition de cette nouvelle génération de progiciels de contrôle des données d'enquête. L'augmentation continue de la puissance des ordinateurs et, simultanément, la diminution du coût des ordinateurs personnels sont les plus importants. Comme le révèlent les travaux de Toro et Chamberlain (1988), il est désormais possible d'effectuer le recensement d'un pays de moins de 10 millions d'habitants sur micro-ordinateurs. Il y a trois ans à peine, la majorité des pays de cette taille se servaient de mini-ordinateurs pour l'entrée comme pour le traitement de leurs données de recensement.

Des progrès majeurs ont également été réalisés dans le domaine du génie des logiciels. Au cours des dernières années, les concepteurs de logiciels ont eu accès à de nouveaux outils. Par exemple, trois des progiciels décrits dans le présent document ont été rédigés en partie ou en totalité en langage de programmation C. Bien que les programmeurs des systèmes UNIX emploient depuis près de 10 ans le langage C, l'arrivée récente de compilateurs utilisant un langage C normalisé pour pratiquement tous les systèmes d'exploitation a fait de ce langage le choix de tous les progiciels professionnels élaborés aujourd'hui. L'efficience et la transférabilité du langage C permettent d'élaborer des progiciels parfaitement adaptés aux systèmes d'exploitation que l'on trouve aujourd'hui sur le marché, et facilement transférables sur les nouveaux systèmes, au fur et à mesure de leur mise au point.

La possibilité d'utiliser une base de données ORACLE sur un SGCI constitue un autre exemple de l'utilisation de nouveaux outils normalisés. Si depuis les années 1970 les systèmes axés sur la recherche peuvent employer l'interface SQL pour accéder à des bases de données relationnelles, il n'y a pas très longtemps que l'on peut accéder avec le même interface SQL à des bases de données commerciales, sur une gamme complète d'architectures d'ordinateur. Ces nouvelles bases de données servent de plate-formes aux concepteurs de logiciels pour la création de nouveaux progiciels transférables à d'autres applications.

Bien que les progiciels de contrôle des données d'enquête disponibles aujourd'hui aient bénéficié des progrès récents de la technologie, il reste encore beaucoup à faire à l'avenir. Nombre des progiciels utilisés aujourd'hui pour le contrôle des données qualitatives, comme CONCOR et ISSA, obligent encore le programmeur à rédiger une description de la procédure à suivre pour le contrôle des données d'enquête. D'autre part, le système CANEDIT de Statistique Canada, qui n'exige comme entrée qu'un ensemble de règles de contrôle, peut

être utilisé par des non-programmeurs. Reste encore à voir si les progrès technologiques permettront de remplacer CONCOR ou ISSA par un progiciel de contrôle efficient et transférable, inspiré de CANEDIT. En outre, on ignore encore si le progiciel SGCI que nous avons évalué dans le présent document aura un rendement suffisamment efficace pour justifier son emploi pour le traitement d'enquêtes de moyenne à grande envergure.

RÉFÉRENCES

Bentley, J.L. (1975). Multidimensional Binary Search Trees Used for Associative Searching. Communications of ACM 18, 9 (Sept. 1975), 509-517.

Bentley, J.L. and Friedman, J.H. (1979). Data Structures for Range Searching. Computing Surveys 11, 4 (Dec. 1979), 397-409.

CENTS (1988). Census Tabulation System. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington, D.C. 20233. PC Release 5.0.

Chernikova, N.V. (1965). Algorithm for Finding a General Formula for the Nonnegative Solution of a System of Linear Inequalities. U.S.S.R. Computational Mathematics and Mathematical Physics, 5, 228-233.

CONCOR (1988). Consistency and Correction Edit and Imputation System. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington. PC Release 3.0.

Cotton, P., Turner, M.J. and Hammond, R. (1979). RAPID: A Database Management System for Statistial Applications. 42nd Biennial Congress of the International Statistical Institue, Manila, December 1979.

Diskin, B. (1986). An Evaluation of Microcomputer Data Entry Software Packages for Use in Developing Countries for Census and Survey Processing. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington.

DATADICT (1987). Data Dictionary System. Computer Software Branch, International Statistical Programs Center, U.S. Bureau of Census, Washington. Release 1.0.

Felligi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of American Statistical Association, 71, 17-35.

GEIS (1988). Generalized Edit and Imputation System, Research and General Systems, Statistics Canada, Tunney's Pasture, Ottawa, Ontario. Release 4.2.

Giles, P. (1988). A Model for Generalized Edit and Imputation of Survey Data. Business Survey Methods Division, Statistics Canada.

Graves, R.B. (1976). CAN-EDIT - A Generalized Edit and Imputation System in a Data Base Environment. The Electronic Data Processing Working Party of the Conference of European Statisticians, Geneva, March 1976 (available from Statistics Canada, Ottawa).

ISSA (1988). Integrated System for Survey Analysis, Institue for Resource Development, Westinghouse, PO Box 866, Columbia, Maryland 21044. Release 1.0.

Kernighan, B.W. and Ritchie, D.M. (1978). The C Programming Language. Prentice-Hall, Inc. Englewood Cliffs, New Jersey 07632.

Kovar, J.G., MacMillan, J.H. and Whitridge, P. (1988). Overview and Strategy for the Generalized Edit and Imputation System. Business Survey Methods Division, Statistics Canada, Ottawa. Working Paper BSMD-88-007E.

ORACLE (1988). Oracle Corporation, Belmont, California. Version 5.1 A.

PCEDIT (1988). Software and Support for Population Data Processing, Department of Technical Co-operation for Development, United Nations, New York, New York 10017. Version 2.1.

REALIA COBOL (1988). REALIA Inc., 10 Riverside Plaza, Chicago, Illinois 60606. Version 3.00.

Sande, G. (1979). Numeric Edit and Imputation. Bulletin of the International Statistical Institue. Proceedings of the 42nd Session, Manila, 455-463.

Toro, V. and Chamberlain, K. (1988). The Use of Microcomputers for Census Processing in Developing Countries: An Update. Symposium '88: The Impact of High Technology on Survey Taking, Ottawa, October, 1988.

LOGICIEL POUR L'ANALYSE D'ENQUÊTE

W. A. FULLER,

Iowa State University

Résumé

Plusieurs techniques analytiques, utilisées pour des échantillons d'enquêtes complexes et nécessitant de nombreux calculs, sont examinées. Au nombre des éléments pour lesquels des programmes ont été mis au point figurent (1) l'estimation de la variance pour des statistiques comme les coefficients de régression, les paramètres des modèles d'erreur de mesure, les données de tableaux à double entrée et les quantiles, et (2) l'estimation pour petites régions à l'aide de composantes de modèles de variance. Un logiciel de modification des données, pour protéger le caractère confidentiel des déclarations individuelles figurant dans les microdonnées diffusées, est en cours d'élaboration.

1. INTRODUCTION

La présente conférence est consacrée aux répercussions de la technologie sur les enquêtes. Je parlerai d'un logiciel informatique qui a été mis au point pour l'analyse des données d'enquête ainsi que d'un autre, en cours d'élaboration, pour la prévention de la divulgation.

Les progrès de l'informatique sont qualifiés de "révolution" par certains. J'estime pour ma part qu'il vaut mieux parler d'évolution parce que le processus s'est déroulé relativement en douceur tout au long de ma carrière statistique. Aujourd'hui, nous pouvons tenir dans le creux de notre main une puissance de calcul supérieure à ce que possédait toute une université il y a 30 ans. Aussi, faisons-nous couramment des calculs qui n'étaient même pas envisageables en 1958.

L'une des branches du calcul, dont il sera question ici, est le calcul de variances et des autres statistiques complexes pour les données des échantillons d'enquêtes.

L'ordinateur pose aussi des problèmes aux statisticiens. Les utilisateurs veulent leurs données sous une forme accessible par machine. Puisque les utilisateurs disposent maintenant de moyens de calcul suffisants pour les grands fichiers de données, ils exigent l'accès aux microdonnées. Ces utilisateurs peuvent effectuer des analyses mais leur matériel et leurs logiciels leur permettent aussi de faire des opérations d'appariement. Cela pose des problèmes au statisticien qui cherche à protéger la confidentialité des données des répondants. Nous travaillons actuellement à un programme qui ajoute de l'erreur aux observations pour réduire le risque d'une divulgation des données.

2. ANALYSE DES DONNÉES

Dans la présente section, nous parlerons de certains logiciels pour ordinateurs personnels mis au point pour des échantillons d'enquête complexes. Le programme de base s'appelle PC CARP et a été conçu pour l'IBM PC. Le programme fut diffusé en décembre 1986. Il s'agit d'une variante du programme pour unité centrale appelé SUPER CARP, élaboré en 1970 par l'Iowa State University.

PC CARP a été conçu pour l'estimation de totaux, ratios, moyennes, quantiles, tableaux à double entrée, coefficients de régression et coefficients de régression logistique, et pour l'estimation de la variance de ces statistiques. Une description de ce programme peut être trouvée ailleurs (voir Schnell et al. (1988)). Par conséquent, je ne ferai que mentionner quelques-unes des propriétés spéciales de PC CARP.

Le programme est conçu pour un IBM PC standard de 450k de mémoire ayant un coprocesseur mathématique. Bien entendu, on peut aussi le faire fonctionner sur les nouvelles machines plus puissantes. Le programme peut traiter un nombre illimité

d'observations ou de strates. PC CARP est piloté par menus. Il incorpore des réponses courtes et des réponses implicites de sorte qu'un opérateur accompli peut toujours l'utiliser rapidement. Le programme produit des messages de diagnostic et comporte les options "aide" et "retour en arrière" pour l'opérateur en difficulté. Tous les calculs de variance sont faits à l'aide d'approximations de Taylor. Voir Fuller (1975) et Binder (1983).

Les différents types d'analyse disponibles sont donnés à la figure 1. Cette dernière figure comprend le menu à partir duquel l'utilisateur choisit le genre de calcul à faire. Les questions au bas de l'écran apparaissent une à une lorsque que l'utilisateur a choisi le type d'analyse. Les questions permettent à l'utilisateur de faire un choix parmi les options disponibles pour cette analyse.

```
                              PC CARP
                        Analysis Specification
 ==================================================================
 Select the number corresponding to the desired type of analysis.  Replace
 // with the chosen analysis number (e.g. 08) and press the ENTER key.

                          AVAILABLE ANALYSES

          1. Totals                      7. Subpopulation Totals
          2. Ratios                      8. Subpopulation Means
          3. Difference of Two           9. Subpopulation Ratios
             Ratios                     10. Subpopulation Proportions
          4. Stratum Totals             11. Two-way Table
          5. Stratum Means              12. Regression
          6. Stratum Proportions        13. Univariate
                                        14. STOP Program Execution

          Key the numeric ID of desired analysis.......01

  Please respond( Y or N) to any additional question(s).
 Press "ENTER" after each response:

 Would you like covariances of the estimates to be computed?.....N
 Would you like the design effect(s) to be computed?.....Y
```

Figure 1 Menu PC CARP pour la spécification de l'analyse

L'option de l'analyse unidimensionnelle permet de produire, entre autres statistiques, des estimations de quantiles et un estimateur de l'erreur-type des quantiles. La première étape du calcul des quantiles est la construction d'un estimateur de la fonction de répartition cumulative et d'un estimé de la variance de la fonction de répartition. Il est possible de calculer les quantiles et les statistiques connexes pour une sous-population en spécifiant une catégorie d'une variable de classification.

Le programme est capable de construire des tableaux à double entrée et de calculer le test d'hypothèse de proportionnalité pour ces tableaux. Le programme produit en option la matrice de covariance des proportions. La construction d'estimations pour des sous-populations équivalant à des tableaux à une, deux, trois et quatre entrées est une option particulièrement utile.

La portion régression du programme donne à l'utilisateur la capacité de calculer des résidus et d'effectuer des tests d'hypothèses pour des sous-ensembles de coefficients. La sortie de base contient les coefficients et les erreurs-types évaluées à partir des approximations de Taylor.

Un supplément a été élaboré pour le PC CARP pour permettre le calcul de variances pour certaines statistiques pour les échantillons stratifiés a posteriori. Les variances s'appliquent

à la stratification a posteriori gamma. Voir Fuller et Sullivan (1987).

PC CARP nécessite un ensemble de données complet pour l'analyse. Ceux qui le désirent peuvent obtenir avec le PC CARP un programme d'imputation "hot deck" appelé PRE CARP. L'opération "hot deck" remplace une valeur manquante par la valeur du même article dans l'enregistrement qui vient immédiatement avant l'enregistrement manquant et qui se trouve dans la même catégorie de variable de classification spécifiée. On peut se servir d'une option spéciale dans le programme pour regrouper une strate d'une unité à la strate adjacente.

PC CARP a été conçu pour des fonctions difficilement exécutables avec d'autres programmes. Par conséquent, PC CARP ne peut, à toutes fins pratiques, servir pour la création de nouvelles variables à partir de l'ensemble de données existant. D'autres programmes, notamment le PC SAS ou le PRODAS, peuvent faire ce genre d'opérations.

Un autre membre de la famille CARP est le programme EV CARP. Les algorithmes dans ce logiciel sont conçus pour l'analyse de régression des données observées sujettes à l'erreur de réponse. Aucun autre logiciel actuellement disponible n'est capable de produire des estimations pour de tels modèles de régression avec des données recueillies par voie d'enquêtes complexes. Le premier algorithme du genre a été conçu par Michael Hidiroglou en 1973. Voir Hidiroglou (1974) et Hidiroglou et al. (1976).

```
                    EV CARP
              Analysis Specification

Select the number corresponding to the desired type of analysis.  Replace
the / with the chosen analysis number and press the ENTER key.

                    AVAILABLE ANALYSES

        0. Weighted Least Squares     3. Minimum Distance (EV3)
        1. Error in Equation (EV1)    4. Instrumental Variables (EV4)
        2. Reliability Ratios (EV2)   9. Stop Program Execution

        Key the numeric ID of desired analysis........3

  Would you like sample summary statistics output (Y or N)?..........Y

  Taylor or Normal(SRS) Variance estimator (T or N)?................T

  Enter the value of the Alpha coefficient adjustment ( 0 or 1 ) ....0

  Would you like Residuals and Predicted values output(Y or N)?......Y

  Would you like to test coefficient sets equal to zero?(Y or N)?....Y
```

Figure 2. Menu EV CARP pour la spécification de l'analyse

Les procédures d'estimation nécessitent de l'information sur la nature de l'erreur de réponse, en plus des observations. L'information sur l'erreur de réponse peut prendre la forme d'estimations des covariances de l'erreur, d'estimations des ratios de fiabilité, ou de variables instrumentales. Ces techniques d'estimation sont étudiées par Fuller (1987).

EV CARP est un programme piloté par menus calqué sur le PC CARP et la première série de menus pour l'introduction des données est essentiellement la même que celle de PC CARP. La figure 2 reproduit l'écran du EV CARP pour la spécification de l'analyse. Les différents types d'analyse correspondent aux différents types d'information utilisables avec de tels modèles. Les questions au bas de l'écran apparaissent une fois le type d'analyse sélectionné et représentent les options rattachées à la procédure sélectionnée.

La sortie pour les analyses disponibles est très semblable à celle d'une régression ordinaire. Les coefficients estimés et les erreurs-types qui y sont associées sont produits. Le programme permet à l'utilisateur d'effectuer des tests d'hypothèses pour des sous-ensembles de coefficients. Les résidus et les valeurs réelles estimées pour les variables sujettes à l'erreur sont disponibles en option avec le programme.

3. PRÉVISION POUR DE PETITES RÉGIONS

Un programme qui calcule des prévisions pour de petites régions a aussi été mis au point par l'Iowa State University. La procédure de prévision est fondée sur les composantes d'un modèle de variance parfois appelé le modèle empirique de Bayes. Le modèle se rapproche beaucoup de certains modèles d'erreur de mesure et cette analogie est exploitée par le programme. Voir Fuller (1986).

Les paramètres du programme sont le vecteur des estimations, lesquelles proviennent de l'enquête pour les régions, toute variable de contrôle pour les régions, et les matrices de covariance des estimations provenant de l'enquête. Dans sa version actuelle, le programme est plutôt petit et le nombre total de variables, la somme du nombre d'estimations et le nombre de variables de contrôle ne doivent pas dépasser dix.

Le programme calcule en deux étapes une approximation de l'estimateur à maximum de vraisemblance de la composante inter-groupes de la variance. Cette matrice de covariances est restreinte de manière à être positive semidéfinie et sert à construire des prévisions pour les petites régions. Un estimateur de la matrice approximative de covariances des prédicteurs est également calculé. Deux formes de la matrice de covariances peuvent être calculées. L'une d'elles est un estimateur de la variance inconditionnelle du prédicteur multidimensionnel. La seconde est un estimateur de la variance conditionnelle du prédicteur, lorsque le conditionnement est sur le vecteur observé faisant l'objet de la prévision.

Le programme est une variante du EV CARP et est piloté par menus de la même façon que les autres membres de la famille CARP.

4. PRÉVENTION DE LA DIVULGATION

En raison du nombre croissant de fichiers de microdonnées qui sont diffusés et des améliorations que l'on continue d'apporter à la théorie du couplage des enregistrements et aux logiciels utilisés pour ce couplage, la confidentialité des réponses fournies par les répondants est de plus en plus menacée.

Pour garantir un degré de confidentialité aux répondants dont les données font partie des microdonnées diffusées, on peut ajouter de l'erreur aux réponses, avant la diffusion. Un programme devant servir à cette fin est actuellement mis au point à l'Iowa State University.

Il est semble-t-il souhaitable que les données diffusées ressemblent le mieux possible à l'échantillon original. Le programme actuel vise à conférer aux données diffusées un vecteur moyen et une matrice de covariances presque égaux au vecteur moyen et à la matrice de covariances des données originales. De plus, toutes les fonctions de répartition de l'échantillon unidimensionnel des données diffusées seront proches des fonctions de répartition correspondantes des données originales.

Le programme procède par étapes. La première étape transforme les observations originales en variables "quasi-normales" à l'aide de l'inverse des fonctions de répartition de l'échantillon. L'erreur normale est ajoutée aux variables transformées, la somme est

standardisée à la variance unitaire et la somme standardisée est transformée à l'échelle originale à l'aide de la fonction de répartition de l'échantillon original. Ce programme est en train d'être rédigé dans le langage matriciel IML de SAS.

REMERCIEMENTS

La mise au point de PC CARP a été rendue possible grace à l'accord 50-YABC-5-66028 avec le International Statistical Programs Center du U.S. Bureau of the Census. Le projet a également profité de l'entente de coopération 68-3A75-8-12 avec le Soil Conservation Service du U.S.Department of Agriculture. La recherche sur la prévention de la divulgation est présentement effectuée en vertu de l'entente statistique conjointe JSA 88-1 avec le U.S. Bureau of the Census.

Les programmes pour ordinateurs personnels PC CARP et EV CARP ont été mis au point par W.J. Kennedy, D. Schnell, G. Sullivan et H.J. Park. Douglas Bearrood et Nancy Hasabelnaby ont participé au développement du programme de prévision pour les petites régions. Gary Sullivan travaille au projet de prévention de la divulgation.

RÉFERENCES

Binder, D.A. (1983), On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

Fuller, W.A. (1975), Regression analysis for sample survey. Sankhya C. 37, 117-132.

Fuller, W.A. (1986), Small area estimation as measurement error problem. In Proceedings of the Conference on Survey Research Methods in Agriculture, (D. Faulkenberry, Ed.) American Statistical Association and NASS, U.S. Department of Agriculture, Washington, D.C.

Fuller, W.A. (1987), Measurement Error Models. Wiley, New York.

Fuller, W.A. and Sullivan, G. (1987), Gamma Post Stratification. Report to the U.S. Bureau of the Census. Dept. of Statistics, Iowa State University, Ames, Iowa.

Hidiroglou, M.A. (1974), Estimation of regression parameters for finite populations. Unpublished Ph.D. thesis, Iowa State University.

Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1976), SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa.

SAS Institute Inc. (1985), SAS Language Guide for Personal Computers Version 6 Edition, Cary, N.C.: SAS Institute Inc.

Schnell, D., Kennedy, W.J., Sullivan, G., Park, H.J., and Fuller, W.A. (1988), Personal computer software for complex surveys. Survey Methodology To appear.

SUDAAN: UN PROGICIEL DÉTAILLÉ POUR L'ANALYSE DES DONNÉES D'ENQUÊTE

L. M. LAVANGE, B. V. SHAH,
B. G. BARNWELL ET J. F. KILLINGER,

Research Triangle Institute

RÉSUMÉ

Le Research Triangle Institute travaille actuellement à l'élaboration d'un progiciel d'analyse des données d'enquête assez souple pour répondre aux besoins des statisticiens d'enquête ainsi qu'à ceux des chercheurs qui ne s'y connaissent pas nécessairement dans le domaine de l'échantillonnage. Ce système a été conçu afin d'apporter des améliorations considérables au progiciel existant du RTI, SUDAAN. Le nouveau système SUDAAN est rédigé en langage C, ce qui le rend plus efficace et plus facilement transférable. La stratégie d'élaboration du logiciel consiste à développer et à mettre au point des procédures statistiques autonomes sur un "PC" d'IBM ayant une capacité de mémoire accrue. Les procédures sont ensuite mises en application dans deux autres contextes informatiques, VAX/VMS et le système d'exploitation IBM. Parmi les nombreuses modifications apportées au nouveau système SUDAAN, on retrouve les options d'estimation de la variance et les techniques d'analyse de données qui n'étaient pas disponibles dans le progiciel RTI. Le présent document décrit la conception et l'élaboration de ce système.

MOTS-CLÉS

Estimation de la variance, interface avec l'utilisateur, langage C.

INTRODUCTION

Les scientifiques du Research Triangle Institute ont consacré les dix-sept dernières années à la conception, à l'élaboration et à la maintenance de systèmes de logiciels pour l'analyse des données d'enquête. Tout récemment, le RTI a été engagé par le Public Health Service (PHS) pour élaborer un progiciel détaillé répondant aux besoins des analystes statisticiens du National Center for Health Statistics (NCHS) et du PHS. Ce progiciel doit être suffisamment souple pour traiter la plupart des plans statistiques utilisés par ces agences. À cette fin, nous avons entrepris d'élaborer un système qui regroupe un grand nombre des caractéristiques du système existant d'analyse de données d'enquête du RTI tout en y apportant des améliorations considérables. Ce document décrit en détail la conception et l'élaboration de ce progiciel détaillé. La conception du système SUDAAN visait premièrement l'élaboration d'une série de procédures permettant d'exécuter une analyse de données statistiques pour les enquêtes par sondage complexes. Le but final était de créer un système qui pourrait contribuer à la conception et à l'évaluation de nouvelles techniques statistiques.

Le système existant de logiciels du RTI consiste en une série de procédures qui produisent des analyses statistiques sous forme de tableaux à double entrée pondérés, des estimateurs généralisés par quotient, ainsi que des régressions linéaires et logistiques. Toutes les procédures ont été constituées à l'aide d'un ensemble unifié de sous-programmes FORTRAN exécutables dans le contexte informatique SAS sur un processeur central IBM. Au cours de l'élaboration du nouveau système, nous avons repensé notre conception actuelle et il en est ressorti quatre objectifs :

- Transférabilité
- Fiabilité/exactitude numérique

- Efficacité de calcul
- Facilité de modification/d'amélioration.

La conception du progiciel détaillé d'analyse de données d'enquête, SUDAAN, a été faite en tenant compte de ces objectifs. On a atteint le premier objectif en rédigeant entièrement le nouveau logiciel en langage de programmation C et en le mettant à l'épreuve simultanément sur plusieurs systèmes, puisque des compilateurs pour le langage C existent pour toute une variété de systèmes d'exploitation. Les programmes initiaux sont actuellement opérationnels avec le compilateur TURBO C sur le "PC" d'IBM, avec le compilateur SVC sur un ordinateur 68020, et avec le système d'exploitation UNIX sur un ordinateur GOULD. Des versions sont également mises à l'essai sur les VAX/VMS et IBM/VS.

On a atteint le deuxième objectif en incorporant des contrôles dans le système de façon à assurer que les définitions des objets ont un sens et que les calculs sont réalisables. Nous avons mis en place des procédures numériques stables pour l'accumulation des sommes de carrés et de produits croisés ainsi que pour les opérations matricelles, y compris l'inversion de matrices.

Pour atteindre le troisième objectif, soit une efficacité accrue, on s'est efforcé lors de la conception d'isoler les tâches qui se répètent très souvent telles qu'un calcul numérique pour chaque variable de chaque document. Ces sections de code sont optimisées dans la mesure du possible sans utiliser le langage d'assemblage. Parmi les algorithmes d'optimisation, on note le traitement des matrices creuses et des vecteurs, le remplacement de la multiplication au moyen de la consultation des tables par l'addition, l'élimination des tests répétitifs "si" (dans la boucle) en créant des boucles distinctes, et bien d'autres éléments.

La création d'un langage de programmation statistique évolué constitue un important produit de l'effort de conception du SUDAAN. Le langage SUDAAN a été développé de façon à faciliter l'accès des statisticiens/programmeurs au progiciel, ce qui a permis d'atteindre le quatrième objectif énoncé ci-dessus. Les procédures statistiques sont actuellement rédigées en langage SUDAAN. Une fois que la version finale aura été rédigée, certaines parties pourront être converties en programmes en langage C afin de maximiser l'efficacité des calculs.

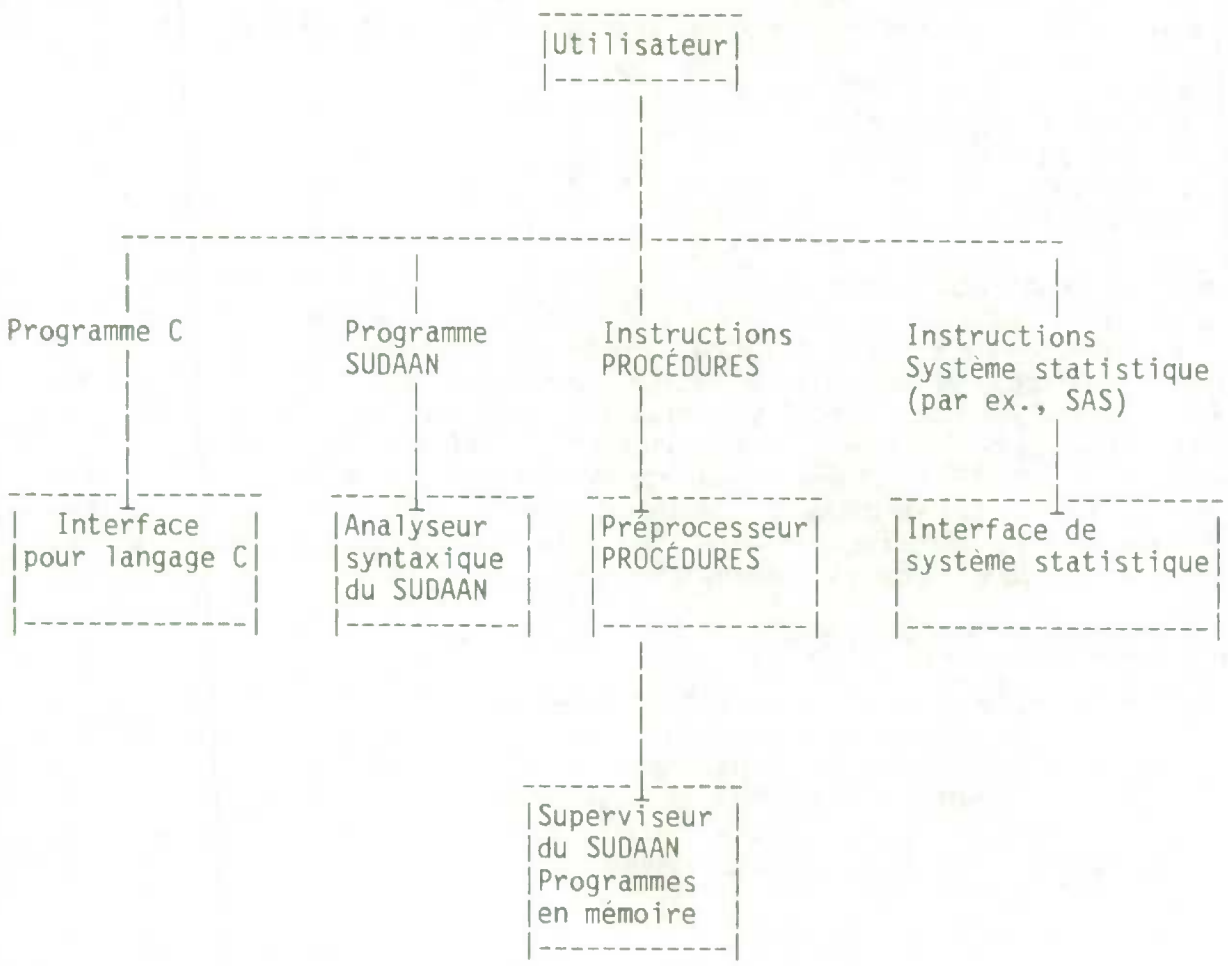
Le plan du SUDAAN prévoit quatre types d'interfaces avec l'utilisateur, tel qu'il est illustré dans le organigramme 1. L'utilisateur général pourra avoir accès au système via des procédures faciles à utiliser. Les procédures seront éventuellement disponibles par l'entremise d'un système statistique, tel que SAS, mais la conception actuelle consiste en procédures autonomes reliées à un format de fichier ASCII standard. L'utilisateur plus expérimenté peut rédiger des programmes SUDAAN pour les calculs lorsque ces produits ne sont pas disponibles dans la bibliothèque de procédures. Cette option d'accès est particulièrement bien adaptée à l'essai de modifications à apporter aux procédures existantes, telles que le calcul de la statistique d'un autre test. Le programmeur en langage C peut accéder directement au système au moyen des fonctions pouvant être appelées.

Les principales caractéristiques du système SUDAAN sont décrites en détail dans les chapitres suivants. La prochaine section donne un bref aperçu du progiciel d'analyse de données d'enquête actuellement mis en marché par le RTI. Elle décrit également les procédures offertes par le nouveau progiciel, avec l'accent sur les différences entre les deux progiciels. La conception du système SUDAAN constitue le prochain sujet . Le langage SUDAAN est illustré dans les exemples de programmation. Les avantages et les inconvénients de la conception du SUDAAN sont exposés à la dernière section.

PROCÉDURES D'ANALYSE DES DONNÉES D'ENQUÊTE DU RTI

HISTORIQUE

Au RTI, la collaboration entre statisticiens et informaticiens, qui a débuté en 1971 (ex., Folsom, Bayless et Shah, 1971) et qui se poursuit toujours, a donné lieu à une série de programmes d'analyse de données d'enquête, lesquels programmes reflètent une méthodologie d'avant-garde dans ce domaine. Tous les progiciels d'analyse du RTI font appel à la série de Taylor ou à la méthode Delta pour estimer la variance dans le cadre des élaborations d'enquêtes par sondage complexes (Folsom, 1974). Chaque programme actuellement mis en marché est conçu pour s'exécuter avec le logiciel SAS.



Organigramme 1. Accès de l'utilisateur au système SUDAAN

Les premiers travaux supposaient une procédure avec des éléments de variance élaborée pour produire des estimations des éléments de la variance associés aux strates et degrés d'un échantillon stratifié par grappes, inégalement pondéré. Une version à jour, VCMPNLS (Shah, 1971) était disponible en 1979 comme procédure SAS. En 1974, le programme STDERR a fourni en premier les estimations pondérées de moyennes, de totaux, de proportions et leurs erreurs type. Par la suite, une autre version, SESUDAAN (Shah, 1981), a permis à l'utilisateur de vérifier s'il y avait des différences entre les estimations par domaine tirées d'une classification à double entrée, et de produire des estimations ajustées à une répartition de population spécifiée par l'utilisateur. Les modifications apportées au SESUDAAN en 1980 et en 1981 ont abouti à deux procédures, RATIOEST (Shah, 1981) et RTIFREQS (Shah, 1982). RATIOEST produit des estimations de quotients de deux variables. RTIFREQS calcule les distributions de fréquences pondérées et les estimations pourcentages de rangées, de colonnes et de pourcentages globaux pour des tableaux à double entrée spécifiques à l'utilisateur. Une autre modification du RATIOEST permettant à l'utilisateur d'introduire un fichier d'unités d'échantillonnage et un ficher d'événements (par ex., visites médicales avec documents multiples par sujet) a mené à la procédure RATIO2 (Shah, 1982).

SURREGR (Holt, 1977), le progiciel de régression d'enquête du RTI, fournit des estimations des coefficients de régression linéaire, de leurs matrices de variance/covariance estimées, et des tests d'hypothèses concernant les coefficients appropriés à une enquête par sondage complexe. L'élaboration du RTILOGIT (Shah et al., 1984) a amené les capacités de régression logistique. Cette procédure s'exécute conjointement avec le PROC LOGIST de SAS et produit des estimations à pseudo vraisemblance maximum des paramètres ainsi que des tests d'hypothèses sur ces paramètres.

Le RTI commercialise actuellement les procédures suivantes pour usage général :

- SESUDAAN
- RATIOEST
- RTIFREQS
- SURREGR
- RTILOGIT

Comme nous l'avons mentionné précédemment, l'estimation de la variance pour toutes ces procédures se fait à l'aide de la méthode de linéarisation par séries de Taylor. Les procédures RTIFREQS, SURREGR et RTILOGIT supposent l'échantillonnage avec remise, ce qui permet d'utiliser une formule de variance de premier degré. Cette méthode est généralement considérée comme une approximation conservatrice pour des variances associées à des plans plus complexes. Les procédures SESUDAAN et RATIOEST permettent à l'utilisateur de travailler avec un échantillonnage avec ou sans remise à chaque degré d'un plan à plusieurs degrés. Les statistiques produites dans le SESUDAAN, le SURREGR et le RTILOGIT reposent sur la statistique de type T^2 de Hotelling, laquelle est censée avoir une distribution asymtotique F transformée (Shah, Holt et Folsom, 1977).

Procédures du nouveau système

La conception du système SUDAAN compte deux étapes d'élaboration de logiciel, chacune menant à une série de procédures statistiques. Les procédures de la première étape touchent principalement l'analyse descriptive de données, y compris les tableaux à double entrée, l'estimation généralisée par quotient et l'estimation des quantiles. Le plan de la deuxième étape, qui est actuellement dans sa phase de planification, englobera plusieurs procédures analytiques, dont la modélisation linéaire et log-linéaire ainsi que l'analyse de survie des données d'enquête.

Le nouveau système SUDAAN offre plusieurs caractéristiques que l'on ne retrouvait pas dans le logiciel du RTI. Les principales sont les multiples options de calcul de variance. Les variances échantillonnales sont calculées pour l'une des trois options du plan demandée par l'utilisateur. Ces options sont les suivantes :

- Échantillonnage avec remise au premier degré
- Échantillonnage aléatoire simple à chaque degré (avec ou sans remplacement)
- Échantillonnage probabiliste inégal sans remise au premier degré et échantillonnage aléatoire simple (avec ou sans remplacement) aux degrés subséquents.

La deuxième option nécessite l'introduction des tailles de population et d'échantillon pour chaque strate et pour chaque degré. Outre ces dénombrements, la troisième option suppose l'introduction des probabilités de sélection conjointe pour chaque paire d'unités primaires d'échantillonnage dans une strate. Ces quantités servent à produire les estimations de la variance Yates-Grundy-Sen pour cette option du plan. Les trois options de calcul de variance couvrent une vaste gamme de plans de l'échantillon incluant l'échantillonnage avec des probabilités proportionnelles à la taille, au premier degré. On peut traiter les probabilités inégales de sélection à un stade subséquent en formant une pseudo strate à ce stade étant donné que les probabilités de sélection peuvent varier d'une strate à l'autre. Outre ces options de variance, l'utilisateur peut demander le calcul de la matrice des variances-covariances pour les estimations à l'intérieur d'une table pour toutes les procédures permettant les tableaux à double entrée. Auparavant, les estimations des matrices des variances-covariances n'étaient disponibles que dans les procédures SURREGR et RTILOGIT.

Le logiciel de la première étape consiste en trois procédures : CROSSTAB, DESCRIPT et RATIO. La procédure CROSSTAB est l'équivalent du RTIFREQS dans le nouveau système SUDAAN. En plus de calculer la fréquence pondérée et les répartitions de pourcentages, elle produit des tests d'indépendance chi carré pour un tableau de contingence à double entrée. Les tests sont évalués à l'aide du paramètre statistique de Wald selon les méthodes préconisées par Koch, Freeman et Freeman (1976). Au cours de la deuxième étape, le paramètre statistique corrigé de Rao-Scott Satterthwaite (Thomas et Rao, 1984 et 1985) sera évalué en sus du paramètre statistique Wald dans le cadre des tests d'ajustement. Thomas et Rao ont démontré que ce paramètre statistique possède les meilleures caractéristiques en ce qui a trait à la puissance et au seuil de signification comparativement à d'autres statistiques qui compensent pour la libéralité du test Wald.

La procédure DESCRIPT calcule les moyennes, les totaux, les proportions, les moyennes géométriques, les quantiles ainsi que leurs erreurs d'échantillonnage. Les estimations peuvent être demandées pour les sous-domaines de tableaux à double entrée spécifiques à l'utilisateur. L'utilisateur a également accès aux options de moyennes et proportions normalisées et(ou) post-stratifiées et à leurs erreurs d'échantillonnage. L'apport des variables de standardisation et(ou) de post-stratification et leurs distributions est requis.

Le RTI vient tout juste d'achever une étude de simulation comparant deux méthodes d'estimation des variances des quantiles pour les données d'enquête de sondage (Wheeless et al., 1988). Il s'agissait (1) de la méthode proposée par Fuller et Francisco (1986) et couramment utilisée dans le progiciel PC CARP et (2) de la méthode de linéarisation directe proposée par le RTI. La simulation comparait également deux méthodes d'estimation des quantiles. Grâce aux résultats de cette étude, la méthode de linéarisation directe est actuellement programmée pour être utilisée dans la procédure DESCRIPT.

L'utilisateur peut demander des estimations des contrastes linéaires parmi les estimations par domaine produites par DESCRIPT. Il y a quatre types de spécifications de contraste disponibles. L'utilisateur peut demander un contraste linéaire général, toutes les paires de

différences possibles parmi les niveaux de variables de domaine, les différences entre les niveaux de variables de domaine et la valeur marginale, et les effets polynomiaux pour une variable de domaine ordinale.

La procédure RATIO calcule les estimations des rapports généralisés pour les données d'enquête. L'utilisateur peut spécifier les variables de numérateurs et de dénominateurs qui sont continues ou discrètes. Dans le dernier cas, on a besoin des niveaux des variables pour lesquels les rapports doivent être estimés. À l'instar de la procédure DESCRIPT, cette procédure peut s'exécuter dans l'un des trois modes suivants : estimation de base, estimation standardisée ou estimation post-stratifiée.

Le plan actuel pour l'élaboration du logiciel de la deuxième étape regroupe quatre procédures statistiques rédigées dans le cadre du nouveau système SUDAAN. La première est une procédure d'analyse de données catégoriques offrant des capacités de modélisation log-linéaire pour l'analyse des données d'enquête. Cette procédure constituera l'analogue de la procédure CATMOD de SAS dans le cas des données d'enquête complexes. Les procédures deux et trois seront des procédures de régression linéaire et logistique, dont l'élaboration repose sur les procédures existantes SURREGR et LOGIST. Elles seront toutes deux reprogrammées dans le nouveau système SUDAAN et plusieurs améliorations y seront apportées. On prévoit des améliorations à la procédure de régression linéaire pour l'analyse de modèle des coefficients de régression stochastique des données d'enquête. Le logiciel a été élaboré par le RTI dans le cadre d'un contrat avec le National Institute for Child Health and Human Development (LaVange et al., 1988) et sera intégré au système SUDAAN. La quatrième procédure, de nature générale destinée à l'analyse des données de survie, devrait permettre à l'analyste d'effectuer une estimation de Kaplan-Meir ainsi qu'une estimation du modèle de risque proportionnel de Cox à l'aide de données d'enquête complexes.

PLAN DU SYSTÈME SUDAAN

L'objectif premier du plan du système SUDAAN consistait à élaborer une série de procédures pouvant effectuer l'analyse de données statistiques pour les enquêtes par sondage complexes. Le but final était d'obtenir un système pouvant contribuer à la conception et à l'évaluation de nouvelles techniques statistiques. Le plan du système SUDAAN est constitué de trois couches :

(1) Des PROC ou procédures similaires aux procédures SAS ou BMDP
(2) Des structures de données SUDAAN et des fonctions opérant d'après ces structures de données
(3) Un interpréteur de niveau du système pour (1) et (2).

Ces couches correspondent à différents aspects du système. En effet, différents groupes d'utilisateurs peuvent avoir des perceptions différentes du SUDAAN. Dans la présente section, nous discuterons de ces aspects et nous décrirons en détail les principaux composants du système SUDAAN.

Perceptions du système

Le chercheur autonome ou l'analyste de données considérera le SUDAAN comme un regroupement de procédures statistiques permettant d'appliquer différentes techniques statistiques pour l'analyse de données d'enquête. Le statisticien le verra comme un outil pratique pour évaluer les nouvelles formules statistiques ou techniques d'analyse. Le SUDAAN prévoit la conversion facile des formules en instructions exécutables, permettant ainsi de programmer rapidement et efficacement de nouvelles procédures ou des améliorations aux procédures existantes.

Le concepteur de système peut le percevoir comme un compilateur, rédigé en langage C, pour un langage de haut niveau afin de préciser des tâches d'analyses statistiques. Pour l'informaticien, l'effort consistait à concevoir un compilateur ou interpréteur optimiseur ayant une capacité d'optimisation globale de toutes les ressources disponibles. Le programme SUDAAN ne définit que ce qui est exécuté et non comment il est exécuté. Le système peut choisir n'importe quel ordre de succession des opérations pour atteindre les résultats définis par l'utilisateur.

L'interface PROC a été intégrée dans le SUDAAN afin d'aider l'utilisateur qui applique sans cesse les mêmes procédures à différents ensembles de données ou différentes variables à l'intérieur d'un même ensemble de données. La perception du programmeur d'étude représente l'une des nombreuses conceptions possibles de compilateur pour la mise en application du SUDAAN.

Toutefois, la principale force sous-jacente à la conception du SUDAAN est la vision du statisticien, à savoir un outil pour traduire les formules mathématiques en un programme machine capable de produire des résultats numériques pour une combinaison de données déterminée. Le statisticien n'a pas besoin de spécifier comment calculer un estimateur particulier. Il se concentre plutôt sur ce qui doit être calculé. Pour lui, le SUDAAN est un langage "non procédural". Vu la définition d'une quantité statistique en termes de formule mathématique, le système détermine l'algorithme de calcul. Les utilisateurs n'ont aucune connaissance relativement à la représentation des structures de données en mémoire ou au dispositif d'entrée/sortie ni aucun contrôle sur le déroulement du programme. En ce sens, le langage du SUDAAN est considéré comme un langage définitionnel et non procédural.

Les éléments de base du système sont les suivants :

- Objets de données statistiques et structures de données abstraites
- Fonctions et opérations par rapport aux structures de données
- Fonctions relatives à l'entrée, à la sortie et à l'impression des structures de données.

Chacun de ces éléments est décrit dans les sections suivantes.

Structures de données statistiques

L'ensemble de données provenant d'une enquête nationale sur les ménages constitue un exemple de structure de données statistiques. Un tel ensemble de données peut être composé de documents comprenant de l'information sur un échantillon d'individus, lesquels sont choisis à partir d'un échantillon de ménages dans des districts, comtés et états de dénombrement. L'ensemble de données réel comprend un dossier pour chaque individu, lequel dossier est constitué de diverses réponses sur les attributs et détails de cet individu et de ce ménage. Chaque document est identifié à l'aide de plusieurs identificateurs hiérarchiques correspondant aux degrés de la sélection de l'échantillon : état, comté, district de dénombrement, ménage et individu. Un ensemble de données récapitulatives sur les dénombrements de population d'un comté déterminé par caractéristiques, telles que la race et le sexe constitue un exemple de type différent de structure de données. Un tel ensemble de données constitue la réunion de tables ou matrices bidimensionnelles et de deux identificateurs hiérarchiques, l'état et le comté.

Dans le SUDAAN, les structures de données statistiques sont définies comme des réunions d'ensembles multidimensionnels avec des identificateurs hiérarchiques emboîtés. On appelle ces structures de données arbres d'ensembles équilibrés (AEE). Chacun des ensembles de

données décrits ci-dessus est un exemple d'AEE, le premier comportant cinq identificateurs emboîtés et le second, seulement deux. Le langage SUDAAN consiste également en une série de fonctions et d'opérateurs mathématiques utilisés pour l'entrée et la sortie de AEE et pour définir les AEE en tant que fonctions de AEE et d'autres objets d'introduction. Le langage SUDAAN est en soi un ensemble de AEE fermé mathématiquement en ce qui a trait à ces fonctions. L'application d'une fonction à un AEE produit un autre AEE. Le niveau d'identificateurs emboîtés peut changer, mais cela produit toujours un ou plusieurs AEE.

Exemples de programme SUDAAN

Pour illustrer les concepts du langage SUDAAN, considérons le problème du calcul des carrés moyens intra-strate entre les unités primaires d'échantillonnage dans le cas de l'estimation du total d'une population. Supposons que y (hijk) désigne la valeur observée de la variable y, recueillie au moyen d'un plan d'échantillonnage emboîté, pour le k-ième individu dans le j-ième ménage, la i-ième unité primaire d'échantillonnage et la h-ième strate. L'indicatif de fichier "EXEMPLE1" renvoie à un fichier ASCII, les identificateurs d'échantillons emboîtés étant compris en tant que variables un à quatre dans le fichier. La variable y est la cinquième variable du fichier. Le fichier d'entrée est stocké dans un format de fichier standard avec un cliché d'article décrivant les variables et les niveaux s'y rattachant. Les étiquettes de variables ou les formats sont automatiquement mis en mémoire et utilisés pour l'impression. Dans les instructions suivantes, X renvoie au AEE composé de ces valeurs de données. Le programme utilise la fonction de totalisation, SIGMA, pour calculer les totaux de y à divers niveaux. Le premier argument de SIGMA renvoie au AEE qui contient les valeurs de données à additionner, tandis que le second argument renvoie au niveau d'emboîtement jusqu'où se fera la totalisation. Cet argument égale le nombre d'identificateurs emboîtés dans le AEE résultant qui renferme les sommes. Les commentaires, délimités par les caractères "/*" et "*/" clarifient davantage la logique des instructions de programme suivantes.

```
RECORDS =   FICSFILE("EXAMPLE1",
            VECTOR(1,2,3,4));          /*RECORDS est un AEE comprenant
                                         l'ensemble de données d'entrée*/

X = SELECT(RECORDS,5);                 /*X est un AEE de valeurs y et
                                         d'identifications*/

PSULEVEL=2;
XSUM2=SIGMA(X,PSULEVEL);               /*Additionner X pour chaque unité
                                         primaire d'échantillonnage*/
MH = SIGMA(CONSTANT(XSUM2,1),1);       /*Calculer le nombre d'unités
                                         primaires d'échantillonnage par
                                         strate en définissant un vecteur
                                         de un pour chaque unité primaire
                                         d'échantillonnage dans XSUM2*/

MHMIN1 = MH - 1;
STRATLEV = 1;
XSUM1 = SIGMA(XSUM2,STRATLEV);         /*Additionner X pour chaque
                                         strate*/
XSQR1 = SIGMA(XSUM2*XSUM2,
              STRATLEV);               /*Additionner par strate les carrés
                                         des totaux d'unités primaires
                                         d'échantillonnage*/
```

```
XSSO = SIGMA((MH * XSQR1 -
        XSUM1*XSUM1)/MHMIN1,0);          /*Calculer la somme des écarts au carré
                                           pour toute la strate*/

MSE = SQRT(XSSO);
PRTABS(MSE);                             /*Imprimer les résultats*/
```

Étant donné que le langage SUDAAN ne fonctionne qu'avec des définitions symboliques, les instructions ci-dessus peuvent facilement être modifiées afin d'obtenir les erreurs quadratiques moyennes associées à plus d'une variable. Dans l'exemple, en remplaçant la seconde instruction par ce qui suit, on obtient le calcul des erreurs quadratiques moyennes pour les instructions 5, 6, 9 et 10.

```
X = SELECT(RECORDS, VECTOR(5,6,9,10));
```

Ensuite, considérons le problème du calcul des erreurs quadratiques moyennes pour les tailles estimées des populations correspondant aux cases d'une table multidimensionnelle. Dans les exemples d'instructions suivants, les TABLES renvoient à un AEE constitué d'une table à double entrée et d'une table à triple entrée. La première est la classification recoupée des variables 5 et 6 du fichier d'entrée et la seconde, la classification recoupée des variables 5, 9 et 10. La fonction EFFECTS convertit les variables catégoriques dotées de numéros spécifiques de niveaux en vecteurs factices (0,1). La fonction CROSSP définit les tables comme des produits croisés de ces vecteurs. SIGMAWG fournit deux arguments, les sommes pondérées et non pondérées pour chaque case des tables. Une fois que XSUM2 est défini, les instructions du programme dans l'exemple ci-dessus entrent en jeu pour calculer les erreurs quadratiques moyennes. Les instructions requises sont :

```
RECORDS = FICSFILE("EXAMPLE1", VECTOR(1,2,3,4));
SUBGROUPS = VECTOR(5,6,9,10);            /*Variables définissant les sous-groupes*/
LEVELS = VECTOR(7,2,2,2);                /*Niveaux de variables associés*/
OPTIONS = VECTOR(0,0,0,0);               /*Les quatre options sont à "0"*/
EFFVECTORS = EFFECTS(RECORDS,SUBGROUPS,LEVELS,OPTIONS);
TAB1 = VECTOR(1,2)                       /*La variable 5 par la variable 6*/
TAB 2 = VECTOR(1,3,4);                   /*La variable 5 par la variable 9 par la
                                           variable 10*/
TABVEC = VECTOR(TAB1,TAB2);
TABLES = CROSSP(EFFVECTORS, TABVEC);
WEIGHT = SELECT(RECORDS,11)              /*La variable 11 constitue le coefficient de
                                           pondération*/
MWSUM = SIGMAWG(TABLES, PSULEVEL, WEIGHT);
XSUM2 = NWSUM(2);
```

On peut ajouter d'autres instructions à ces exemples pour calculer les erreurs quadratiques moyennes des pourcentages estimés pour chaque table. Ceux-ci nécessitent des fonctions SUDAAN supplémentaires et sont illustrés dans l'exemple de programme donné à l'annexe A. Ce programme illustre l'estimation de la variance pour la procédure CROSSTAB en supposant un échantillonnage au hasard simple et stratifié avec remise.

Procédures SUDAAN

L'élaboration de nouvelles procédures dans le SUDAAN est une tâche assez simple. Le procédé suppose la rédaction d'un programme SUDAAN et l'association des objets du programme à un ensemble bien défini de mots-clés. L'entrée de l'utilisateur pour une PROC

consiste en une syntaxe de niveau de commande, comme celle qui est utilisée dans le IBM/TSO ou SAS. Le compilateur PROC traduit les données entrées par l'utilisateur et génère les instructions nécessaires pour l'analyseur syntaxique du SUDAAN. Bref, la rédaction d'une procédure SUDAAN équivaut à définir un préprocesseur pour un programme SUDAAN.

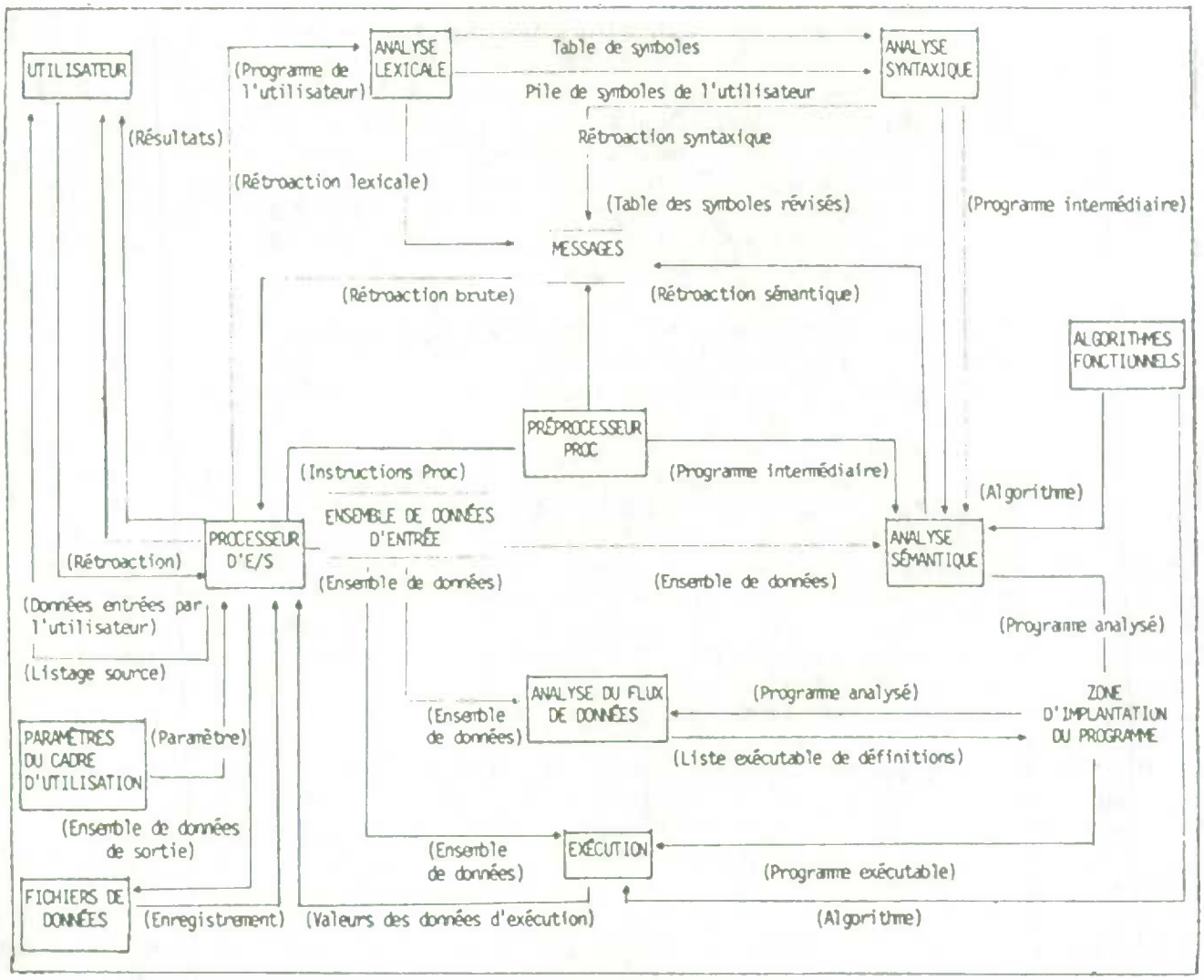
Le principal avantage de cette méthode réside dans le fait que l'utilisateur dispose d'un programme SUDAAN pour chaque procédure utilisée comme documentation détaillée. L'utilisateur peut voir la traduction des données entrées en résultats de même que la série de calculs effectués une fois que la procédure a été appelée. On peut apporter des modifications à la procédure en variant le programme SUDAAN s'y rattachant sans rédiger une procédure entièrement nouvelle. La source de toutes les procédures SUDAAN sera mise à la disposition de l'utilisateur. La source de la procédure CROSSTAB, illustrée à l'annexe A, en constitue un exemple.

Composants du système

Le système est constitué de six principaux composants :

- analyseur lexical
- analyseur syntaxique
- analyseur sémantique
- superviseur d'exécution
- définitions de fonctions
- compilateur PROC.

L'organigramme 2 illustre le déroulement global du système. Comme nous l'avons déjà mentionné, tout le système est rédigé en langage de programmation C. Le programme fourni par l'utilisateur passe d'abord dans l'analyseur lexical qui génère une table de symboles. L'analyseur syntaxique traite cette table afin de produire une liste d'instructions exécutables. L'analyseur sémantique vérifie la validité des types de fonctions, les arguments compatibles ainsi que l'interprétation significative des instructions du programme. S'il y a lieu, à ce stade, les messages d'erreurs sont produits. Une étape intermédiaire de sortie permet de stocker le programme analysé si on le veut. Un programmateur optimiseur réorganise l'ordre d'exécution des instruments du programme aux fins d'utilisation optimale des ressources. Le superviseur d'exécution produit ensuite les résultats du programme. Le processeur d'E/S fournit au système l'interface d'utilisateur adéquate. Cela peut comprendre une interface de procédures qui transforme les instructions de traitement en une série d'arguments pour le programme SUDAAN.



Organigramme 2. Conception du système SUDAAN

OBSERVATIONS FINALES

La caractéristique qui distingue le SUDAAN est l'entité désignée sous le nom de AEE. Sur le plan mathématique, l'AEE représente une séquence finie d'ensembles multidimensionnels. La série d'AEE représente un ensemble fermé en ce qui a trait aux opérations et aux fonctions binaires dont les arguments sont des AEE. La syntaxe fonctionnelle est une fonction "naturelle" pour l'expert en statistique mathématique qui peut utiliser le langage SUDAAN pour définir les AEE qui l'intéressent.

Le code source SUDAAN fait partie de la documentation mise à la disposition de l'utilisateur du SUDAAN. En se reportant au programme SUDAAN pour exécuter une procédure statistique, l'analyste sera en mesure de déterminer exactement quelles sont les formules mathématiques fondamentales employées. Lorsque la recherche statistique signale des formules nouvelles ou améliorées pour une analyse donnée, les formules peuvent être facilement traduites en langage SUDAAN pour produire un programme machine de travail.

L'utilisateur qui rédige un programme SUDAAN ne fait que définir les objets de calcul au moyen des cases de fonction. Il n'a pas à préciser comment les calculs seront effectués. On obtient une fiabilité accrue grâce aux contrôles intégrés dans le système lesquels permettent d'assurer des définitions significatives et la génération des objets requis pour la sortie. Le système est libre de choisir l'algorithme de calcul optimal étant donné l'ensemble de données et les ressources de calculs disponibles. Le système offre également énormément de souplesse d'emploi en ce qui a trait à l'entrée et à la sortie. Tous les objets d'entrée/sortie sont définis comme AEE. Tout AEE intermédiaire peut être sauvegardé ou affiché à n'importe quel moment dans le programme.

Malgré les avantages énumérés ci-dessus, l'utilisation des AEE comporte certaines restrictions. En effet, l'utilisateur doit se limiter aux fonctions d'AEE actuellement en application et tous les calculs doivent être représentés comme des AEE ou des fonctions de ceux-ci.

Le programmeur SUDAAN doit s'en tenir au vocabulaire de fonctions disponible. Les calculs qui nécessitent de nouvelles fonctions d'AEE peuvent devoir être reportés jusqu'à ce que de telles fonctions soient mises en application dans le SUDAAN. Avec le temps, un nombre croissant de fonctions seront mises en application et cette restriction ne sera pas aussi critique. La conception modulaire du système permet d'y ajouter de nouvelles fonctions sans recompiler tout le système. Seules la nouvelle fonction et son interface doivent être compilées. Il est également possible de tester la nouvelle fonction de façon indépendante avant de l'intégrer au système, ce qui facilite grandement l'ajout de fonctions à la bibliothèque du système, au fur et à mesure des besoins.

Le SUDAAN ne pourra pas effectuer les calculs présentés sous forme d'AEE. L'utilisateur doit définir tous les calculs comme des fonctions d'AEE d'entrée ou des données d'exécution. Par conséquent, le système SUDAAN n'est pas destiné à la programmation générale, mais devrait s'avérer adéquat pour l'analyse statistique et tout particulièrement pour l'analyse de données d'enquête.

Presque tous les efforts d'élaboration de logiciel au RTI ont été orientés vers la conception du système SUDAAN. Quoique cette stratégie ait initialement ralenti l'élaboration des procédures statistiques, avec le temps, nous prévoyons des économies considérables en raison de la facilité avec laquelle chaque procédure peut être conçue, mise en application et mise au point. Ces efforts produiront non seulement un progiciel facile à utiliser qui s'exécute efficacement, mais également une aide précieuse au statisticien pour l'amélioration des procédures existantes et la planification de procédures supplémentaires au fur et à mesure que la nouvelle méthodologie sera disponible.

REFERENCES

COX, D.R. (1972). "Regression Models and Life Tables" (with discussion), Journal of the Royal Statistical Society B, 34: 187-220.

FOLSOM, R.E., BAYLESS, D.L., and SHAH, B.V. (1971). "Jackknifing for Variance Components in Complex Sample Survey Designs." Proceedings of the American Statistical Association: 36-39.

FOLSOM, R.E. (1974). "National Assessment Approach to Sampling Error Estimation. Sampling Error Monograph. Prepared for the National Assessment of Educational Progress.

FRANCISCO, C.A. and FULLER, W.A. (1986). "Estimation of the Distribution Functon with Data from a Complex Survey." Presented at the American Statistical Association Meetings, Chicago, Illinois.

HOLT, M.M. (1977). SURREGR: Standard Errors of Regression Coefficients from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC 27709.

KAPLAN, E.L. and MEIER, P. (1958). "Nonparametric Estimation from Incomplete Observations." J. of the American Statistical Association, 53: 457-481.

KOCH, G.G., FREEMAN, D.H., JR. and FREEMAN, J.L. (1975). "Strategies in the Multivariate Analysis of Data from Complex Surveys." Inter. Stat. Rev. 43, 59-78.

LAVANGE, L.M., PFEFFERMANN, D., SHAH, B.V., and GABEL, T.J. (1988). "The Analysis of Relationships Between Growth and Dietary Intake Using the NHANES Dataset". Final Project Report. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B.V., HOLT, M.M. and FOLSOM, R.E. (1977). "Inference about Regression Models from Sample Survey Data." Proceedings of the 3rd Meeting, International Association of Survey Statisticians, New Delhi.

SHAH, B.V. (1981). RATIOEST: Standard Errors Program for Computing Ratio Estimates from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B.V. (1982). RTIFREQS: Program to Compute Weighted Frequencies, Percentages, and Their Standard Errors. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B.V. (1981). SESUDAAN: Standard Errors Program for Computing Standardized Rates from Sample Survey Data. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B.V. (1978). "SUDAAN: Survey Data Analysis Software." Presented at the Joint Statistical Meetings of the American Statistical Association, San Diego.

SHAH, B.V. (1979). VCMPNLS: Program to Compute Variance Components. Research Triangle Institute, Research Triangle Park, NC 27709.

SHAH, B.V., FOLSOM, R.E. HARRELL, F.E. and DILLARD, C.N. (1984): "RTILOGIT: Survey Data Analysis Software for Logistic Regression", Final Report, Work Assignment 74, Research Triangle Institute, Research Triangle Park, NC.

THOMAS, D.R. and RAO, J.N.K. (1984). "A Monte Carlo Study of Exact Levels of Goodness-of-Fit Statistics Under Cluster Sampling." Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C.

THOMAS, D.R. and RAO, J.N.K. (1985). "On the Power of Some Goodness-of-Fit Tests Under Cluster Sampling." Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, DC.

WHEELESS, S. and SHAH, B.V. (1982). "RATIO2 and RATIO3: Computng Ratio Estimates with Two Data Files. Research Triangle Institute, Research Triangle Park, NC.

WHEELESS, S.C., SHAH, B.V., and LAVANGE, L.M. (1987). "Results of a Simulaton for Comparing Two Methods for Estimating Quantiles and their Varinces." Project Report. Research Triangle Institute, Research Triangle Park, NC 27709.

ANNEXE A

Exemple de programme SUDAAN

```
/* Exemple d'appel au PROC CROSSTAB */

PROC CROSSTAB DATA=test120 DESIGN=wrs;
NEST stratum psu segno sampno;
SUBGROUP agecat race dsmsa dpovind;
LEVEL 7 2 2 2;
WEIGHT lwt;
TABLES race*dpovind agecat*dpovind agecat*dsmsa*dpovind;
```

C R O S S T A B

Instructions du programme SUDAAN

```
1  #include "constant.def"          /*fichier définissant les constantes*/

2  filename = "test120";            /*fichier de données d'entrée*/

3  nestvar=vector(1,2,3,4);         /*position de la strate, de l'unité primaire
                                      d'échantillonnage, du segment et de l'identifi-
                                      cation de la personne sur l'enregistrement
                                      d'entrée*/

4  wgtvar=5;                        /*position de la variable du coefficient de
                                      pondération sur l'enregistrement des données
                                      d'entrée*/

5  desnam = "WRS";                  /*l'option de conception égale l'échantillonnage
                                      de remplacement*/

6  psulev = 2;                      /*position de la variable d'unité primaire
                                      d'échantillonnage sur le vecteur d'instruction
                                      emboîté*/

7  strlev = 1;                      /*position de la variable de strate sur le
                                      vecteur d'instruction emboîté*/

8  subgroups=vector(8, 9, 10, 11);  /*positions de l'ÂGE, de la RACE, du "SMSA",
                                      de la STRATE DE PAUVRETÉ sur
                                      l'enregistrement d'entrée*/
```

```
9  levels = vector(7, 2, 2, 2);          /*niveaux pour chaque variable de sous-
                                           groupe*/

10 optionvec=vector(0,0,0,0);            /*paramètres de l'option : Inclure, Afficher,
                                           Nomarg et Mergehi;*/
                                         /*1="Oui", 0="Non"*/
                                         /*définitions des tables à l'aide des positions
                                           des variables dans le vecteur de SOUS-
                                           GROUPE*/

11 tab1 = vector(2, 4);                  /*RACE par STRATE DE PAUVRETÉ*/

12 tab2 = vector(1, 4);                  /*ÂGE par STRATE DE PAUVRETÉ*/

13 tab3 = vector(1, 3, 4);               /*ÂGE par "SMSA" par STRATE DE
                                           PAUVRETÉ*/

14 tables = vector(tab1, tab2, tab3);    /*AEE combinant trois tables pour le traite-
                                           ment*/

15 records=ficsfile(filename),           /*introduire les enregistrements avec les
           nestvar, wgtvar);             variables emboîtées et pondérées s'y
                                           rattachant*/

16 effvectors = effects(records          /*créer (0,1) les variables d'indicateur pour
                 subgroups, levels,      les niveaux de sous-groupes ou les vecteurs
                 optionvec);             d'effets*/

17 tabvec=crossp(effvectors,tables);     /*former les produits croisés des vecteurs
                                           d'effets*

18 w = select(records, wgtvar);          /*introduire les coefficients de pondération*/

19 nsum0 = sigma (tabvec, 0);            /*calculer les dénombrements
                                           fragmentaires*/

20 wsum0 = sigma(w*tabvec, 0);           /*calculer les dénombrements pondérés*/

21 rowper=margop(wsum0,wsum0,ROW,        /*calculer les pourcentages par rangée*/
           "/");

22 toomplement = margop(tabvec,
           tabvec, ROW, "-");

23 z=w*(tabvec*(1-rowper) +              /*niveau individuel de linéarisation des séries
          tcomplement * rowper)            de Taylor*/
         / wsum0;

24 zpsu = sigma(z, psulev);              /*somme par unité primaire d'échantillonnage
                                           des écarts aux séries de Taylor*/

25 zstr = sigma(zpsu, strlev);           /*somme par strate des écarts aux séries de
                                           Taylor*/

26 z2str = sigma(zpsu * zpsu, strlev);   /*somme par strate des écarts quadratiques*/
```

```
27 npsu = sigma(constant(zpsu, 1),
        strlev);                          /*nombre d'unités primaires d'échantillonnage
                                            prélevées à l'intérieur d'une strate*/

28 varwr = sigma((npsu*z2str-zstr
          * zstr)/(npsu-1), 0);           /*avec variance de remplacement du
                                            pourcentage par rangée*/

29 serow = sqrt(varwr);                   /*erreur standard de pourcentage par
                                            rangée*/

30 print(vector(rowper, serow),
      vector("Row %", "S.E. of row
      %"));                               /*impressions du % par rangée et du % par
                                            rangée d'erreur standard pour les 3 tables
                                            définies aux lignes 11 à 13*/
```

C R O S S T A B

Formules statistiques pour l'exemple de programme SUDAAN

1 h = strate, $h=1,2,\ldots H$

i = unité primaire d'échantillonnage; $i=1,2,\ldots n_h$

j = unité d'échantillonnage; $j=1,2,\ldots n_{hi}$

w_{hij} = coefficient de pondération par individual (hij)

r = rangée de table; $r=1,2,\ldots R$

c = colonne de table, $c=1,2,\ldots C$

P_{rc} = pourcentage de rangée pour la case rc

23

$$z_{hij}(rc) = \begin{cases} 0 & \text{si l'unité d'échantillonnage hij n'est pas dans la rangée r} \\ w_{hij} * (-p_{rc'}) & \text{si l'unité d'échantillonnage hij est dans la rangée r et dans la colonne } c' \neq c \\ w_{hij} * (1 - p_{rc}) & \text{si l'unité d'échantillonnage hij est dans la case rc} \end{cases}$$

24 $z_{hi} = \sum_j z_{hij}$

25 $\sum_i z_{hi}$

26 $\sum_i z^2{}_{hi}$

27 n_h

28 $Var = \sum_{h=1}^{H} \{ n_h \sum_i z_{hi}^2 - (\sum_i z_{hi})^2 \} / (n_h - 1)$

29 $Serow = \sqrt{Var}$

CONTRIBUTIONS DE L'INTELLIGENCE ARTIFICIELLE (IA) AU TRAITEMENT DES DONNÉES STATISTIQUES

J. H. JOHNSON,

Statistique Canada

Contributions de l'Intelligence Artificielle (IA) au Traitement des Données Statistiques

J. H. Johnson

Statistique Canada
Sous-division de la recherche
et des systèmes généraux
Pièce 2306, Edifice principal
Ottawa, K1A 0T6
(Original in English)

Résumé

La recherche dans le domaine de l'intelligence artificielle (IA) a permis de faire beaucoup de progrès dans l'exploitation de la technologie informatique. Certains de ces progrès peuvent être appliqués à l'acquisition, au traitement et à l'analyse de l'information effectués par un organisme statistique comme Statistique Canada. On peut compter au nombre de ces contributions, l'utilisation de langages de programmation très évolués, d'environnements de programmation de l'IA, de la notion de la représentation des connaissances, de meilleures interfaces homme-machine, du raisonnement superficiel, du raisonnement approximatif, du traitement d'un langage naturel, du poste de travail de l'IA, des systèmes experts et des coquilles de systèmes experts. Après avoir présenté chacune de ces contributions, nous décrirons brièvement comment elles peuvent être mises en application dans le domaine du traitement des données statistiques.

MOTS CLÉS: *Intelligence Artificielle; Applications; Traitement des Données Statistiques; Systèmes Experts.*

La recherche dans le domaine de l'intelligence artificielle (IA) [3,6,8,13] a permis de faire beaucoup de progrès dans l'exploitation de la technologie informatique. Des notions comme le calcul symbolique (la manipulation de l'information symbolique par opposition à l'information numérique) et la récurrence, considérée comme une technique de programmation, tirent leur origine des premiers travaux en IA et comptent maintenant parmi les outils de travail de chaque programmeur. Des travaux plus récents en IA visent également à proposer des techniques, des méthodes, des outils et des idées qui seront aussi très vraisemblablement acceptés dans la pratique générale de la science informatique.

La personne qui n'est pas familière avec ces nouvelles notions se retrouve malheureusement plongée dans les ténèbres de l'IA. Ceux qui commentent les possibilités de la technologie font des déclarations que la "sagesse conventionnelle" du professionnel de l'informatique ne semble pas pouvoir appuyer. De plus, l'enthousiasme des praticiens et des chercheurs en IA concernant leurs "jouets" et les possibilités de ceux-ci, ne rassure pas les personnes qui sont plus intéressées à voir à ce que le travail soit fait.

A quels avantages pouvons-nous donc nous attendre de la part de cette technologie de l'IA? Nous concentrerons notre examen sur les besoins d'un organisme statistique, comme Statistique Canada, en matière de traitement et de gestion des données. Le terme traitement des données statistiques (TDS) correspondra à l'opération de traitement des données qui vient appuyer les opérations d'acquisition, de traitement et d'analyse de l'information ainsi que d'accès à l'information recueillie

par un tel organisme. Un grand nombre d'observations auront, naturellement, une applicabilité plus grande, mais cela n'est pas intentionnel.

Le reste de ce document donnera un aperçu général de ce qui constituera, selon l'auteur, les principaux progrès en recherche en IA dont pourra bénéficier le TDS. Ces contributions comprennent l'utilisation

1. des langages de programmation très évolués (LPTE),
2. d'environnements de programmation de l'IA,
3. de la notion de représentation des connaissances (RC),
4. de meilleures interfaces homme-machine (IHM),
5. du raisonnement superficiel,
6. du raisonnement approximatif,
7. du traitement d'un langage naturel,
8. du poste de travail de l'IA,
9. des systèmes experts (SE), et
10. des coquilles de systèmes experts (CSE)

Après avoir présenté chacune de ces contributions, nous décrirons brièvement comment elles peuvent être mises en application.

Vous trouverez tout au long des pages qui suivent le terme "système général". Celui-ci désigne un système informatique qui appuie un aspect quelconque du processus de TDS sans être étroitement lié à une enquête particulière. Il comprend habituellement une méthodologie qui a évolué dans le temps suite à l'application fructueuse de techniques à des enquêtes particulières, ainsi que des méthodes plus complexes nécessitant plus de ressources qu'une seule enquête ne peut justifier. Ainsi, le système général CANEDIT peut effectuer une vérification et une imputation des performances en se fondant seulement sur la définition du fichier d'entrée et sur des "règles de vérification" évoluées.

Langages de Programmation Très Évolués

On appelle souvent les langages de programmation utilisés dans les travaux en IA, langages très évolués (LPTE). Le dynamisme de la recherche dans le domaine des LPTE n'est pas seulement conditionné par les besoins de la communauté de l'IA, mais aussi par la volonté plus générale d'élaborer des langages de programmation qui soient faciles à utiliser tout en étant efficaces, notamment, dans les architectures en parallèle. Trois méthodes principales sont examinées: la programmation fonctionnelle, la programmation logique et la programmation orientée objets.

Jusqu'ici, on se servait des langages LISP[11]. et APL ainsi que du langage de Backus en *programmation fonctionnelle*. Plusieurs langages ont vu le jour récemment suite aux efforts visant à créer un langage facile à utiliser exempt de tout effet secondaire, et qui s'appuie sur la théorie du lambda calcul (typé). L'un des plus connus d'entre eux est le ML[14]. L'avantage de ces langages est que les programmes qui en sont issus sont très faciles à lire et à rédiger et qu'ils peuvent être traduits sous une forme exécutable très efficace. Ils réunissent certaines des meilleures caractéristiques du APL (abstractions évoluées proches de la conceptualisation du programmeur) ainsi que la sémantique claire du lambda calcul et la lisibilité d'un langage conçu pour être utilisé.

La *programmation logique* tire son origine du succès qu'a connu le Prolog [7,12]. En se fondant sur la méthode de démonstration d'un théorème par résolution, elle considère le calcul comme le moyen de démontrer la validité d'un prédicat en refutant sa négation. Le calcul des résultats se fait à l'aide de liages introduits par le procédé d'unification, c'est-à-dire, en se servant des implications des restrictions imposées aux valeurs pendant le calcul. Les nouveaux travaux visent à développer le modèle tout en conservant l'efficacité de base du Prolog standard.

La *programmation orientée objets* relie le calcul aux objets sur lesquels une opération est effectuée. Pour trouver, par exemple, quelle valeur est associée à la clé d'accès 346 dans une structure de données de type "dictionnaire", on envoie le message `at: 346` à la structure de données. Celle-ci se sert de sa méthode `at:` pour extraire la valeur. La nature de l'algorithme utilisée se situe donc au niveau de l'objet de la structure de données. Les

objets sont introduits dans une hiérarchie des héritages propres aux méthodes. Par conséquent, la méthode correspondant à `at:` sera fort probablement rattachée à une classe qui est un ancêtre véritable de la classe de la structure de données elle-même.

Un autre exemple suppose l'affichage ou l'impresion des objets. Chaque objet peut accepter un message "print", auquel cas son contenu est affiché d'une certaine façon à l'intention de l'utilisateur. Ainsi, les connaissances relatives à la forme de l'impression sont décentralisées (se rapportent aux objets eux-mêmes) plutôt que centralisées (groupées en un programme d'impression gigantesque comportant une importante instruction de cas).

Vous trouverez dans un appendice un examen plus détaillé des LPTE ainsi qu'un exemple donnant une comparaison des styles.

LPTE et TDS

Les langages très évolués offrent aux programmeurs la possibilité d'améliorer leur rendement en leur permettant d'utiliser des abstractions plus évoluées et des méthodes de programmation déclarative. Ils inciteront peut-être plus les utilisateurs finals à rédiger eux-mêmes des programmes, en leur faisant sauter les étapes d'analyse et d'élaboration de système et en leur ouvrant la voie vers une méthode plus expérimentale du traitement des enquêtes.

A certains égards, toutefois, il s'agit là d'une tendance qui se dessinait depuis un certain temps et qui ne fait que se poursuivre. Les systèmes généraux tendent à produire naturellement leurs propres ordres de gestion dont la complexité n'a d'égale que la fonctionalité. Un système qui contrôle la vérification et l'imputation doit être régi par des règles de vérification qui peuvent être très facilement spécifiées à l'aide d'un langage déclaratif conçu à cette fin. De plus, l'imputation nécessite la définition des conditions selon lesquelles un enregistrement donneur convient à une imputation suivant la méthode hot-deck, ainsi que de la façon de procéder pour faire l'imputation. Le passage à un LPTE pour effectuer l'adaptation nécessaire, par opposition à l'utilisation d'un langage de programmation conventionnel comme le PL/1, ne représente plus qu'une simple étape supplémentaire.

Environnements de Programmation de l'Intelligence Artificielle

Les travaux dans le domaine de l'IA font appel à une très grande diversité d'environnements de programmation. Toutefois, les environnements articulés autour du LISP et du Prolog (ou de certains dérivés) sont les plus populaires. Le Smalltalk, bien que sa création ne soit pas liée à l'IA, a été utilisé avec succès pour les travaux en IA.

Ces trois langages de programmation considèrent le calcul comme une opération fondamentalement interactive et, en conséquence, les environnements comprennent habituellement des outils qui facilitent l'élaboration interactive d'applications interactives. De plus, tous les langages renferment la fonction de liage tardif qui permet de différer le plus tard possible l'attribution de la signification des noms. Puisque l'efficacité d'un compilateur dépend d'un liage précoce, un grand nombre de langages de l'IA doivent, compte tenu de leur nature même, être interprétés plutôt que compilés. Si les langages sont compilés, le code résultant prend de l'expansion en raison de tous les contrôles d'exécution effectués, ou devient simplement une succession d'appels de sous-programme.

Le besoin d'efficacité a été établi. Cela a entraîné la création de compilateurs pour le langage LISP qui, en réaménageant quelque peu la sémantique et en interdisant certaines techniques de programmation, peuvent produire un bon code machine. Il existe des compilateurs de code naturel pour le Prolog qui donnent d'excellents résultats. La compilation du Prolog est plus réalisable parce que ce langage est plus récent et que la sémantique est plus simple. Les personnes chargées de la mise en application du Smalltalk ont incorporé certains liages précoces dans des cas spéciaux pour que la compilation puisse produire un meilleur code machine. Ainsi, les véritables liages associés aux objets True et False sont ignorés pour permettre l'exécution d'essais immédiats des conditions.

Ces environnements de programmation comprennent habituellement une importante bibliothèque de programmes pour prendre en charge une IHM de haut niveau au moyen d'un terminal en mode point et autoriser un accès de haut niveau aux fichiers, ainsi que pour assurer la vérification

en direct du code source ou des données, et d'autres fonctions, le cas échéant.

Bien que la disponibilité des outils d'élaboration et l'efficacité du système livré aient une incidence sur le choix de l'environnement, le langage utilisé constitue également un facteur.

LISP est un vieux langage qui a produit de nombreux dialectes. Par conséquent, le logiciel peut difficilement être transféré d'un système LISP à l'autre à moins d'utiliser un sous-ensemble très restreint. Les efforts récents visant à uniformiser les dialectes pour former un dialecte commun qui porte un nom bien choisi, le COMMON LISP [11], ont permis de surmonter quelque peu cette difficulté. Beaucoup de fournisseurs du LISP soutiennent maintenant qu'ils vendent une version de ce langage uniformisé ou, du moins, qu'ils travaillent à obtenir la compatibilité des dialectes.

Le Prolog représente une contribution beaucoup plus récente à la famille des langages de programmation. Conçu par Colmerauer et Roussel qui cherchait à créer un langage à partir de la méthode de démonstration de théorème par résolution de Robinson, il a rapidement fait sa marque comme langage de programmation complet. Parce qu'il s'apparente à la logique clausale et à l'inférence, qui se fonde sur le principe du chaînage arrière, il est utilisé par beaucoup de chercheurs en IA qui considèrent la logique clausale comme une bonne technique de représentation des connaissances. Le Prolog a des avantages sur le LISP en ce sens que l'unification et le retour en arrière font partie intégrante du premier, tandis que le second doit faire appel à l'utilisateur qui mettra ces caractéristiques en application et qui, pour cette raison, sera souvent pénalisé pour s'être servi d'un moins bon outil. Un Prolog offrant une "optimisation récursive" et un bon programme récupérateur peut émuler un grand nombre des caractéristiques utiles du LISP.

Le Smalltalk-80, bien qu'il soit utilisé sous une forme ou sous une autre depuis dix ans, n'est recommandé que depuis peu à titre de langage de l'IA. Il représente l'exemple type des langages que l'on désigne sous le nom de "langage orienté objets". Le Smalltalk-80 intéresse les spécialistes de l'IA en raison de sa ressemblance avec la technique des "cadres" de la représentation des connaissances en IA. Ces cadres, appelés aussi systèmes de "remplissage de tiroirs", organisent l'information en hiérarchies où le cadre enfant prend implicitement les valeurs de son parent à moins qu'une valeur de substitution ne soit explicitement donnée.

Environnements de Programmation de l'IA et TDS

Les langages et les environnements de programmation de l'IA conviennent-ils au TDS? Les perspectives d'établissement d'un créneau sont bonnes, mais ces langages et ces environnements ne remplaceront pas les méthodes "conventionnelles" propres à la plupart des logiciels puisque leur coût d'exécution est actuellement très élevé. À plus long terme, grâce à une réduction du coût des ressources, ils gagneront probablement en popularité lorsque les systèmes experts articulés autour d'eux seront mis sur le marché. Ils seront également présents dans les applications qui n'exigent pas beaucoup de ressources informatiques, comme c'est le cas des systèmes frontaux raccordés aux systèmes généraux où la partie calcul gourmande en données pourrait être mise en oeuvre dans un langage conventionnel qui accède aux données par l'intermédiaire d'un système de gestion de base de données (SGBD).

Notion de la Représentation des Connaissances

Une autre notion importante qui est à la base de la recherche en IA est celle de la représentation explicite des connaissances inhérentes à une application. Cette notion contraste nettement avec la pratique conventionnelle où l'information est dispersée dans différents endroits et, dans la plupart des cas, sous une forme non assimilable par une machine et non écrite.

Pour éclaircir encore plus ce point, prenez l'exemple d'un système d'application d'une taille appréciable (comme un code source de plus de 20,000 lignes) et posez-vous les questions suivantes:

1. Où sont les connaissances indiquant comment le système est organisé et comment il a été conçu, c'est-à-dire comment il doit être mis en application?

2. Où sont les connaissances indiquant ce qui doit être réellement mis en application?

3. Où sont les connaissances nécessaires à l'exécution du système?

4. Où sont les connaissances requises pour établir les paramètres qui permettront au système de donner un rendement optimum?

5. Si le système se comporte bizarrement, est-il possible d'obtenir une explication des conclusions auxquelles le système en est arrivé et, dans l'affirmative, où sont les connaissances qui décrivent ce raisonnement?

Voici quelles seront les réponses à chacune de ces questions pour beaucoup de systèmes:

1. Les grandes lignes de l'organisation et de la conception du système sont exposées dans un document de spécifications et (ou) un document conceptuel présenté(s) ou non sous une forme assimilable par une machine. Si le document est établi sous une forme assimilable par une machine ce n'est que pour des fins de mise en forme du texte; la machine ne peut pas comprendre le document ni commenter intelligemment le contenu de celuici. La plupart des renseignements importants n'ont peut-être pas été écrits parce qu'ils étaient alors généralement compris ou jugés insignifiants. L'équipe de conception/mise en application originale constitue l'unique source de ces connaissances, bien que certains renseignements puissent être déduits au moyen d'un désossage du code source.

2. L'arbitre final qui décide de ce qui a vraiement été mis en application est le programme lui-même rédigé en code binaire ou en code source. Les observations mêmes du code source n'indiquent peut-être pas correctement ce qui a été mis en application, bien qu'elles renseignent sur les buts visés ou donnent des indications sur la façon d'examiner une structure de données ou une partie de code pour en comprendre la signification.

3. Les connaissances nécessaires à l'exécution du système sont probablement formulées dans le guide de l'utilisateur, bien que celui-ci soit peut-être incomplet, périmé ou difficile à comprendre. Dans le cas de tous les systèmes importants, les connaissances relatives au fonctionnement du système se trouvent, en réalité, dans la tête des personnes qui font fonctionner les systèmes. Ces personnes connaissent les conditions environnementales spéciales nécessaires au bon fonctionnement des systèmes ainsi que les fonctions qui posent des problèmes qu'il faut contourner.

4. Les connaissances concernant la façon d'établir les paramètres pour obtenir un rendement optimum sont distribuées dans toute la communauté des utilisateurs. Si le système est populaire et suffisamment complexe, les utilisateurs peuvent se regrouper pour partager leurs connaissances et même aller jusqu'à organiser des conférences annuelles où ils pourront examiner de nouveaux moyens de se servir du système. Plusieurs "gourous" y révéleront peut-être la bonne façon d'établir les paramètres nécessaires à la réalisation des buts visés.

5. Les connaissances requises pour obtenir du système les raisons de son comportement se trouvent dans le programme lui-même (code binaire ou code source). Bien que ce programme soit compréhensible par une machine en ce sens que la suite des opérations est spécifiée pour que la machine puisse les exécuter, la seule avenue qui s'offre en réponse à la question "pourquoi" est un affichage des instructions exécutées fourni par un logiciel de mise au point universel ou une trace du fonctionnement d'un logiciel de plus haut niveau articulé autour d'une copie de contrôle du code source. Pour interpréter ces renseignements, il faut avoir une connaissance détaillée des mécanismes internes du système et, à cette fin, consulter la réponse à la question 2.

La signification du mot "connaissance" devrait maintenant être plus claire, du moins pour une personne qui travaille en IA. Il s'agit d'une représentation symbolique qui donne suffisamment de renseignements pour que ceux-ci puissent être utilisés pour définir une opération quelconque. La notion d'opération est importante, car elle met l'accent sur une activité plus étendue que la mémorisation d'une certaine structure de surface. Elle s'apparente à l'image de l'instituteur au primaire qui dirait "Décrivez, dans vos propres mots, le message de l'auteur" ou qui ferait faire des exercices qui exigeraient que l'élève considère un problème d'une autre façon que celle qui est présentée.

Par exemple, une base de données stocke des données, c'est-à-dire des chiffres et du texte représentés dans une configuration binaire. Les don-

nées interprétées par un utilisateur ou un analyste deviennent des informations. Le nombre 25,309,330 en lui-même est une donnée, mais devient une information lorsqu'il représente la population du Canada [10]. Une description de la façon dont on arrive à ce nombre à partir d'autre nombres, comme c'est le cas pour une feuille de calcul électronique, serait un exemple de connaissance.

Pourquoi cette notion de la connaissance est-elle importante en IA? On peut répondre à cette question en observant qu'un aspect important de l'intelligence provient de l'aptitude à comprendre et à expliquer le comportement et le raisonnement de quelqu'un. Une intelligence artificielle doit donc être en mesure de représenter et de manipuler explicitement une partie importante de ses connaissances. La compréhension des langages naturels est aussi confrontée à ce problème philosophique: qu'entend-on par "la machine comprend certaines choses?" La meilleure réponse consiste à dire que le texte a été traduit sous une forme qui renferme exactement l'information implicite du texte et qui peut néanmoins être manipulée assez directement par la machine pour faire des déductions (inférences) ou expliquer des faits sous-entendus.

Le système Smalltalk constitue un exemple intéressant de la façon adoptée par l'IA pour envisager les problèmes 1 à 5 qui ont été examinés ci-dessus.

1. L'organisation du système Smalltalk [4,5] fait partie de la structure de la base de données qui renferme les programmes source et objet (byte code) et qui est présentée à l'utilisateur au moyen de l'interface du lecteur. Cette organisation suit une hiérarchie à cinq (5) niveaux [5]: catégorie de classe, nom de classe, méthode d'instance ou de classe, catégorie de message et sélecteur de message. Le lecteur se sert d'une interface articulée autour des concepts de pointage, de défilement et de cliquage qui permet à l'utilisateur de choisir un article à chaque niveau. Après que l'utilisateur a fait une sélection, la fenêtre qui apparaît ensuite propose un menu d'options pour le niveau suivant. Le dernier niveau permet de choisir la fonction à afficher et une sélection à ce moment-ci entraîne l'affichage du programme source dans une fenêtre. Certains niveaux sont clairement conçus à des fins organisationnels et renferment certaines observations. D'autres fournissent d'abondantes observations.

2. Lorsqu'un utilisateur choisit un article source à l'aide des moyens susmentionnés, il obtient le programme source courant, soit le programme source du système qui est alors exécuté. Par conséquent, l'information relative à tous les aspects du système est directement accessible à l'utilisateur en tout temps. Le programme source est entièrement rédigé en langage Smalltalk-80 qui est un langage très évolué.

3. Les connaissances de base nécessaires au fonctionnement du système se présentent sous la forme d'un document imprimé, bien que des écrans d'aide puissent facilement être créés. Après avoir appris quelques principes fondamentaux, l'utilisateur peut facilement se servir du lecteur pour obtenir des connaissances plus spécialisées.

4. La méthode Smalltalk permet de créer des outils très généraux qui seront communiqués à tous les utilisateurs au sein des groupes d'utilisateurs Smalltalk.

5. Le système peut être conçu pour expliquer son comportement au moyen de traces de haut niveau, ou en modifiant directement le programme source et en observant les effets. Bien que dangereuse (une image non viable peut être produite), cette méthode offre une grande flexibilité. Il est également possible de visualiser directement l'état d'un objet à l'aide d'un inspecteur.

Voici les principales techniques de représentation des connaissances:

- Logique: Depuis Aristote, les hommes ont essayé de formaliser la langue et le raisonnement dans le but de clarifier le processus de la pensée. Il en a résulté un système complexe (logique des prédicats) capable de décrire beaucoup d'aspects du monde. Les connaissances peuvent être représentées de manière déclarative en affirmant que plusieurs formules sont vraies. Celles-ci s'appellent habituellement des axiomes. On dit pour toute formule qui est vraie chaque fois que l'axiome l'est, qu'elle est la suite logique de l'axiome. Ce système est irréfutable dans la mesure où l'on peut démontrer par une série de petits

raisonnements valables toutes les formules qui découlent logiquement des axiomes, et toutes celles qui n'en découlent pas.

- Cadres: Une autre notion d'organisation des connaissances qui remonte aux philosophes grecs est celle d'un système de classification hiérarchique. Il s'agit de la notion qui permet de classer les livres dans une bibliothèque selon la matière, puis selon le domaine dans la matière, le sous-domaine dans le domaine, et ainsi de suite. Dans le contexte de l'IA, un cadre est une série de "tiroirs" qui peuvent tous recevoir une valeur ou prendre une valeur implicite stockée dans une mémoire ancestrale au sein de la hiérarchie. Il y a différents systèmes qui offrent plusieurs mécanismes pour remplir les tiroirs de valeurs implicites.

- Réseaux sémantiques: Un réseau sémantique renferme une série de noeuds désignant des concepts et des arcs qui représentent les liens entre les noeuds. La principale différence entre cette méthode et celle des cadres est que le réseau sémantique n'a pas besoin d'une structure hiérarchisée.

- Règles: Une façon populaire de représenter les connaissances dans les systèmes experts fait appel à des règles. Il s'agit des conditions examinées sous la rubrique "raisonnement superficiel" ci-dessous et qui sont étroitement liées aux énoncés de logique clausale et, par voie de conséquence, à la programmation logique.

RC et TDS

La perspective offerte par l'introduction des techniques de représentation des connaissances dans le domaine du TDS est plutôt prometteuse. En plus des applications des systèmes experts qui seront examinées un peu plus loin, l'enregistrement plus détaillé des connaissances portant, par exemple, sur la réalisation d'une enquête, sur un plan général ou particulier, constituera un apport au chapitre de la gestion de l'enquête et simplifiera l'utilisation des systèmes généraux. D'une façon précise, un système général qui comprend la structure de la base et la conception d'une enquête donnée sera en mesure de faire de meilleures recommandations concernant l'estimation, la vérification et l'imputation ainsi que les méthodes à suivre pour protéger la confidentialité des données, qu'un système qui ne possède pas cette information.

Meilleures Interfaces Homme-Machine

Une autre caractéristique que l'on associe souvent à l'IA à l'extérieur des laboratoires de recherche, est le souci de la qualité des interfaces homme-machine. Ce souci découle en partie de la complexité des systèmes eux-mêmes et du besoin qui s'ensuit de contrôler leur comportement, ainsi que de la création de systèmes conçus à l'intention des personnes qui ne sont pas familières avec les ordinateurs. Un autre aspect à considérer concerne la nature inhérente de l'intelligence elle-même qui suppose également la communication des idées au moyen, par exemple, d'un langage naturel.

Le document Apple Desktop Interface donne un bon exemple des principes qui sous-tendent une bonne conception d'interface. Il commence par cette citation:

> «The Apple Desktop Interface is the result of a great deal of concern with the human part of human-computer interaction. It has been designed explicitly to enhance the effectiveness of people. This approach has frequently been labeled *user friendly*, though *user centered* is probably more appropriate.»

Vient ensuite une description des dix principes de conception fondamentaux.

1. «Metaphors from the real world» c'est-à-dire, le système se sert d'une image que l'utilisateur connait bien et avec laquelle celui-ci se sent à l'aise. Dans le cas d'Apple, il s'agit d'un ordinateur de table et de chemises que l'on peut déplacer et dans lesquelles on peut mettre et extraire certaines choses.

2. «Direct manipulation» c'est-à-dire, les gestes posés par l'utilisateur devraient provoquer des effets observables pour donner à celui-ci l'impression de maîtriser la situation.

3. «See-and-point (instead of remember-and-type)» c'est-à-dire, le système ne devrait pas

être conçu pour obliger l'utilisateur à se souvenir de quelque chose que l'ordinateur sait déjà. Il est de beaucoup préférable de se fier à l'aptitude de l'utilisateur à reconnaître ce qui est affiché. Les menus, par exemple, ont été établis à partir de ce principe, mais le document Apple Desktop Interface va beaucoup plus loin que la simple utilisation de menus.

4. «Consistency» c'est-à-dire, les composantes qui se comportent identiquement devraient être présentées de la même façon à l'utilisateur, à l'intérieur d'une application ou d'une application à l'autre. Cela suppose, entre autres choses, l'établissement de normes pour les IHM.

5. «WYSIWYG (what you see is what you get)» c'est-à-dire, l'utilisateur ne devrait pas avoir à se souvenir de secrets, de trucs ou d'informations cachées. Il devrait pourvoir constater les effets réels des gestes qu'il pose.

6. «User control» c'est-à-dire, l'utilisateur devrait avoir l'impression de maîtriser la situation. Il ne devrait pas, notamment, sentir qu'il est obligé d'attendre que le système lui demande quelque chose, mais savoir qu'il peut proposer lui-même, le cas échéant, des informations.

7. «Feedback and dialog» soit le principe selon lequel l'ordinateur indique ce qu'il fait à l'utilisateur et communique dans la langue de ce dernier.

8. «Forgiveness» c'est-à-dire, le système devrait accepter que l'utilisateur puisse faire des erreurs et laisser libre cours à l'expérimentation. Si l'utilisateur fait une opération irréversible, l'ordinateur devrait indiquer la nature du danger et permettre l'annulation de l'opération.

9. «Perceived stability» c'est-à-dire, les images-écran devraient toujours se ressembler et respecter l'emplacement relatif des éléments présentés.

10. «Aesthetic integrity» c'est-à-dire, la présentation sur l'écran devrait être attrayante.

Le document Apple Human Interface Guidelines [1] décrit ces principes et la façon dont ils se rattachent à cette interface particulière. Il donne notamment les détails des modalités à suivre au chapitre des applications Apple pour que les personnes chargées de la mise en oeuvre de celles-ci soient uniformes dans leur travail. Les utilisateurs croient qu'il est important de respecter les détails de cette interface; par exemple, les analyses de logiciel évaluent le degré de conformité à la norme comme une mesure de la qualité.

Les interfaces en langage naturel contribuent également à offrir de meilleures IHM; elle seront examinées un peu plus loin.

Amélioration des IHM et TDS

Pourquoi voulons-nous de meilleures interfaces avec le logiciel? Simplement parce que ce sont les personnes qui utilisent les programmes et qu'elles sont plus productives si les programmes sont faciles à utiliser. Ainsi, les coûts de formation peuvent être réduits dans le cas d'un système qui peut enseigner les concepts et les commandes au fur et à mesure des besoins. De plus, le temps perdu parce qu'un utilisateur ne se souvient pas d'une commande requise ainsi que les coûts d'opportunité encourus parce qu'un système n'est pas du tout utilisé, peuvent être réduits.

Raisonnement Superficiel

Un principe important dans beaucoup de travaux en IA comporte la notion de "raisonnement superficiel". Veuillez noter que ce terme n'a pas de connotation négative. Il désigne simplement le raisonnement abstrait de haut niveau exercé par les humains pour être en mesure de bien comprendre le monde qui les entoure. Le raisonnement profond, par contre, correspond au type de calcul qui est plus à la portée des ordinateurs que des humains, soit beaucoup de calculs faisant appel à des algorithmes spécialisés.

D'une façon générale, le raisonnement est superficiel s'il se fonde sur des règles empiriques, et il est profond s'il se fonde sur des "principes fondamentaux" relativement sûrs et sur le calcul des implications de ceux-ci dans un cas donné. Le terme "superficiel," comme beaucoup d'autres en informatique, est relatif: le raisonnement peut être plus ou moins superficiel.

Considérons, par exemple, un programme qui joue aux échecs ou au bridge. Un système axé sur un raisonnement profond s'appuierait principale-

ment sur les règles du jeu et fouillerait implicitement dans le grand arbre des ramifications futures possibles du jeu, c'est-à-dire l'arbre présentant la position en cours considérée comme la racine, les positions possibles après le coup suivant une étape plus loin, les positions possibles en réponse aux coups possibles de l'adversaire encore une étape plus loin, et ainsi de suite.

Un système axé sur un raisonnement superficiel jouerait à la façon des humains. Un programme d'échecs déplacerait les pièces de manière à attaquer systématiquement une partie de la défense adverse et évaluerait les coups en fonction de leur effet sur l'espace, sur les facteurs de mobilité et sur la valeur des pièces prises des deux côtés. Un programme de bridge ferait ses enchères en se fondant sur le compte de points.

La connaissance superficielle (c'est-à-dire, la connaissance utilisée dans le raisonnement superficiel) se présente habituellement sous la forme de règles énoncées comme suit:

```
IF Condition_1
AND Condition_2
AND Condition_3
AND ...
AND Condition_n
THEN Conclusion
```

Les conditions identifient la situation qui permet d'appliquer la "règle empirique", et établissent des "bornes" qui limitent encore plus l'application de la règle à des domaines où sa véracité est plus certaine. Une autre forme de borne, qui sera examinée dans la section suivante, concerne le problème des règles empiriques qui sont très rarement universellement vraies.

Les calculs qui se fondent sur un raisonnement profond peuvent toujours prendre la forme de règles, mais cela n'est pas souvent le cas puisque la fine structure du problème suggère habituellement des optimisations qui se calculent plus directement si on envisage le problème d'une autre façon. La recherche faite par un programme d'échecs, par exemple, peut être optimisée à l'aide de l'algorithme alpha-beta, et les améliorations massives sont le fruit de l'utilisation des instructions booléennes et de décalage du matériel sous-jacent pour produire des coups autorisés.

Il est intéressant de constater que les meilleurs programmes de jeu s'appuient, en grande partie, sur le raisonnement profond. La régularité et les possibilités d'optimisation offertes par les systèmes axés sur le raisonnement profond semblent l'emporter sur le coût d'une recherche force brute, ce qui a désappointé beaucoup de chercheurs en IA qui croient qu'en donnant plus de connaissances à un programme, celui-ci sera meilleur qu'un programme non intelligent qui considère les aspects, bons ou mauvais.

Le raisonnement superficiel, pour sa part, triomphe lorsque les "principes fondamentaux" ne sont pas évidents, trop éloignés du problème à résoudre, ou mal connus. C'est le cas du diagnostic médical qui sert à identifier une maladie à partir de ses symptômes. À l'intérieur de domaines d'intérêt relativement "restreints", les systèmes qui s'appuient sur le raisonnement superficiel ont pu imiter les aptitudes des spécialistes à faire des diagnostics, entraînant la notion de "compétences" au sens d'ensemble de connaissances, surtout superficielles, acquises par l'expérience, et celle de "système expert" considéré comme un système informatique qui saisit ces compétences et les connaissances qui permettent de les mettre en application. Cette "méta-connaissance", ou connaissance de la façon d'utiliser les connaissances de base, est importante parce qu'elle contribue à une recherche fructueuse des possibilités et donne l'impression que la recherche est orientée vers l'humain qui se sert du système. Vous trouverez un peu plus loin une analyse des systèmes experts.

Ce qu'il convient de souligner ici est que la décision d'opter pour un raisonnement superficiel ou profond aura des conséquences sérieuses sur les performances du système obtenu. S'il est applicable, le raisonnement profond devrait être pris en considération.

Connaissance Superficielle et TDS

La connaissance superficielle n'est pas un domaine nouveau pour le TDS puisque les règles de vérification, les techniques d'imputation ainsi que les règles et les poids des liaisons d'articles représentent tous une forme quelconque de connaissance superficielle. L'expérience qui découle de la recherche en IA au chapitre de l'exploitation de la connaissance superficielle peut être très utile à un environnement qui fait déjà appel à beaucoup de connaissances à base de règles.

Raisonnement Approximatif

Comme nous l'avons déjà mentionné, les règles empiriques ont souvent besoin de "bornes" parce qu'elles ne sont pas absolument sûres. Elles ne sont qu'approximativement vraies ou que généralement vraies.

Cette caractéristique a été observée dans l'élaboration de systèmes de diagnostic médical et a entraîné l'incorporation du raisonnement approximatif dans le paradigme des systèmes experts. Il s'agit, fondamentalement, d'introduire des degrés de certitude. Ainsi, la valeur "un" peut indiquer la certitude, le zéro absolu, la fausseté, et les valeurs intermédiaires peuvent désigner les niveaux intermédiaires de croyance en la véracité et la fausseté de la question. Ou encore, un "un" peut signifier la certitude et un "zéro", une absence complète de connaissances, et les valeurs intermédiaires peuvent indiquer des niveaux intermédiaires de croyance. Dans chaque système, chaque fait sera associé à une valeur et l'application de chaque règle entrainera un ajustement des valeurs qui tiendra compte des degrés de croyance véhiculés par la règle.

Il y a quatre approches fondamentales (comportant un grand nombre de variations):

- Facteurs de certitude: Il s'agit d'une méthode spéciale qui a été incorporée dans le système expert de diagnostic médical MYCIN pour offrir un raisonnement approximatif. La valeur "un" indique la certitude, le zéro, une absence de connaissances, et la valeur "moins un", la certitude quant à la fausseté du fait. Les règles comportent aussi des facteurs de certitude, et un système de formules permet de combiner les valeurs qui produisent un facteur de certitude. Cette approche se prête très facilement à la programmation et ne requiert pas beaucoup de calculs supplémentaires, mais ne repose pas sur une théorie solide. La signification des nombres n'est pas très claire.

- Théorie des ensembles flous: La notion de raisonnement flou part de l'idée que dans le monde réel les faits et les règles sont vrais jusqu'à un certain point. Cette notion conduit à celle de l'indexation des états du monde selon une échelle de nombres entre zéro et un indiquant le degré de véracité. Les faits et les règles correspondant à la valeur "un" sont absolument (toujours) vrais. Les faits et les règles dont la valeur est supérieure ou égale à 0.9 seront vrais pour tous les modèles ayant une valeur inférieure ou égale à 0.9. D'une façon générale, les faits et les règles dont la valeur est supérieure ou égale à p seront vrais pour tous les modèles ayant une valeur inférieure ou égale à p. Les modèles flous s'appuient donc sur une base théorique et sur une sémantique claires, n'ont pas besoin de beaucoup de nombres, sont faciles à calculer et demeurent stables après des applications répétées de la règle. Le principal problème de ces modèles est qu'ils ne correspondent pas nécessaire ment à ce que l'utilisateur pense que les nombres devraient signifier.

- Inférence bayésienne: L'approche bayésienne attribue un nombre entre zéro (fausseté) et un (certitude) à chaque fait. Ce nombre est interprété comme une probabilité subjective et manipulé au moyen du calcul des probabilités. Les nombres sont attribués aux règles comme des probabilités conditionnelles, et sont combinés à l'aide de la règle de Bayes. Cette approche repose sur une solide théorie et est facile à comprendre. Malheureusement, elle ne tient pas compte de la notion d'ignorance et n'offre aucun moyen d'indiquer une absence de connaissances. De plus, elle est coûteuse et nécessite l'établissement de plusieurs paramètres qui, pour une large part, peuvent être très difficiles à évaluer correctement.

- Théorie de l'évidence de Dempster-Shafer: Cette approche [9] généralise l'approche bayésienne pour tenir compte du problème de l'expression de l'absence de connaissances. Dans le cas où il y a deux hypothèses, deux valeurs sont exprimées: p_t, la croyance subjective en la véracité du fait, et p_f, la croyance subjective en la fausseté du fait. Il n'est pas nécessaire que $p_t + p_f = 1$, bien que la somme doivent être inférieure ou égale à un. La différence entre 1 et $p_t + p_f$ indique l'absence de connaissances et $p_t + p_f = 0$ indique un manque total de connaissances. La méthode Dempster-Shafer fait appel à un plus grand nombre de paramètres que la méthode bayésienne, quoique certains problèmes posés par cette dernière soient résolus par la possibilité offerte par la première d'exprimer une ignorance totale ou partielle.

Raisonnement Approximatif et TDS

L'utilisation de la connaissance approximative est implicite dans l'utilisation d'une enquête-échantillon et l'identification des erreurs d'échantillonnage et d'observation. Le principal avantage du raisonnement approximatif consisterait à simplifier l'incorporation de cette connaissance partielle dans le traitement et l'analyse qui se fondent sur celle-ci. Le modèle de liaison des articles de Fellegi-Sunter, par exemple, se fonde sur un modèle de vérification des hypothèses qui se rapproche de la méthode subjective bayésienne.

Une suggestion plus radicale d'application du raisonnement approximatif consisterait à utiliser des règles de vérification qui ne signaleraient pas une erreur absolue, mais mettraient en lumière des conditions inhabituelles ou suspectes. S'il devenait évident qu'une zone ou qu'un ensemble de zones soient peu vraisemblables, une imputation pourrait être requise. L'utilisation de règles de vérification permettrait de détecter des anomalies dans les données d'enquête sans enfermer celles-ci dans des contraintes trop rigides.

Traitement d'un Langage Naturel

Une composante très importante dans les annales de la recherche en IA est le traitement d'un langage naturel. Cette recherche a porté sur la génération et la compréhension des langages naturels ainsi que sur la traduction automatique. Toutes ces applications, si elles sont menées à bien, pourraient améliorer sensiblement l'accessibilité aux systèmes informatiques en permettant une interaction en langage naturel. De plus, un vaste ensemble de connaissances en langage naturel assimilable par la machine ouvriraient de nouvelles avenues: les ordinateurs seraient en mesure d'assimiler de grandes quantités d'informations, qu'ils filtreraient et résumeraient pour les personnes qui aimeraient être informées de tout ce qui se passe dans notre société moderne en rapide évolution, mais qui n'en ont pas le temps. La traduction automatique permettrait aussi de briser les barrières linguistiques sans forcer personne à parler anglais.

Malheureusement, comme il reste beaucoup de travail à accomplir, ce ne sont pour l'instant que des rêves futuristes. Il semble, par exemple, que pour éliminer correctement l'ambiguité de la plupart des textes, beaucoup de connaissances sur le monde réel en général et sur le domaine à l'étude soient nécessaires. Une solution globale passerait donc probablement par la solution de certains problèmes d'IA plus fondamentaux et plus difficiles.

Quelles sont les capacités actuelles des ordinateurs? Les ordinateurs ont été utilisés avec succès dans des domaines d'intérêt restreint, particulièrement ceux qui font appel à un vocabulaire spécialisé où la syntaxe est limitée et la sémantique, précise. Certains trouvent étrange que les ordinateurs comprennent plus facilement un acte authentique, comme un testament, qu'un roman, par exemple.

Les systèmes qui utilisent des interfaces en langage naturel doivent aussi faire face à un sérieux problème. Étant donné que l'ordinateur ne comprend pas nécessairement toutes les informations données par l'opérateur, il peut en tirer de fausses conclusions. Il faudrait prévoir des mesures de protection contre les mauvaises conséquences que cela pourrait entraîner. Par exemple, l'organisation des dialogues entre l'ordinateur et l'opérateur devrait être ainsi faite que l'information déduite par le premier pourrait être vérifié par le second. De plus, la vérification doit se faire naturellement pour que l'opérateur ne soit pas tenté de sauter cette opération. L'information doit avoir été extraite et traitée. Une simple transformation syntaxique de l'entrée n'est probablement pas suffisante pour détecter la plupart des problèmes sémantiques.

L'opérateur doit également connaître les limites du fragment de langage naturel que l'ordinateur comprend, c'est-à-dire, apprendre d'une manière ou d'une autre à communiquer en se servant de termes que le programme connaît. Ainsi, un jeu informatique peut obliger l'utilisateur à se servir du verbe prendre, au sens d'acquérir un objet, et à taper "prendre sceptre". Il ne comprendra peut-être pas la commande "ramasser sceptre". Veuillez noter que le simple fait d'ajouter "ramasser" comme synonyme de "prendre" ne résout que temporairement le problème. Et qu'advient-il des termes "acquérir", "saisir", "empoigner", "obtenir", "soulever", etc...? On peut remédier à cette situation en notant sur une page une liste des verbes de base permis ainsi que des indications permettant d'obtenir plus de détails sur l'utilisation de verbes précis, comme "prendre".

Il n'en demeure pas moins que les programmes qu'offrent une interface en lagage naturel sont plus faciles à utiliser. Les gens préfèrent communiquer avec les ordinateurs en se servant d'un sous-langage bien défini d'un langage naturel plutôt que d'un langage spécialisé conçu par un analyste fonctionnel pour lui permettre d'accomplir une tâche.

TLN et TDS

La contribution du traitement d'un langage naturel au TDS se manifeste de deux façons. La première, que l'examen qui précède laisse supposer, s'insère dans le cadre d'une meilleure interface homme-machine avec les systèmes généraux. Une autre forme d'application concerne l'acquisition des données elles-mêmes. La plupart des données recueillies sont exprimées très naturellement par le répondant à l'aide de quelques mots ou d'une phrase. Pour autant que ces mots sont simples, une consultation de table peut suffire après une correction des fautes d'orthographe. Toutefois, les technologies de traitement d'un langage naturel peuvent s'avérer utiles si les textes sont plus longs. Dans la mesure où le programme comprend une faction suffisante des textes entrés et qu'il sait quand il ne comprend pas (de manière à pouvoir consulter un expert humain), il peut être rentable de saisir le texte en entier et de l'analyser plus tard, plutôt que de le faire analyser à même le questionnaire par des employés de bureau et de saisir seulement le code attribué par ceux-ci. Il en résultera probablement une plus grande uniformité pour l'ensemble de l'enquête et une meilleure qualité de la vérification rétrospective des informations provenant de l'enquête.

Le Poste de Travail de l'IA

Les besoins particuliers de la recherche en IA ont eu pour conséquence naturelle l'amélioration du logiciel et du matériel de soutien. En ce qui concerne le matériel, cela s'est traduit par des machines à hautes performances qui peuvent, notamment, prendre en charge les exigences énormes des programmes LISP importants. Ces machines ont besoin d'une mémoire centrale et d'une mémoire auxiliaire d'une grande capacité, et doivent être très rapides.

Pourquoi les programmes LISP ont-ils besoin d'autant de ressources pour obtenir de bonnes performances? Il faut chercher dans l'indifférence face à l'optimisation de l'utilisation de la mémoire pour les structures de données, le besoin de faire appel à une mémoire centrale de grande capacité. Beaucoup de pointeurs sont utilisés et ceux-ci doivent être stockés. Une liste LISP nécessite au moins deux pointeurs pour chacun de ses éléments: un pour désigner l'élément et un pour désigner le reste de la liste. L'utilisation de structures de données qui conviennent mieux à d'autres langages peut également nécessiter l'emploi de pointeurs supplémentaires. En valorisant l'affectation d'un nouvel espace mémoire plutôt que la réutilisation circonspecte de celui-ci et le recours à un programme récupérateur pour réclamer de la place, nul doute qu'une mémoire de grande capacité ne soit nécessaire pour éviter d'avoir à récupérer très souvent de l'espace.

Le besoin d'une vitesse plus grande part également du raisonnement ci-dessus. Il faut suivre les pointeurs pour trouver les éléments d'information et cette démarche doit être faite chaque fois que la valeur est requise. De la même façon, la récupération d'un espace mémoire peut nécessiter beaucoup de temps puisqu'elle doit "vérifier" la mémoire pour déterminer si le programme en suspens peut potentiellement atteindre chaque bloc.

Une autre raison qui explique le besoin d'une plus grande puissance est que la méthode d'élaboration des logiciels de l'IA fait appel au prototypage qui utilise des algorithmes inefficaces. Cette méthode permet au programmeur de se concentrer sur des idées plutôt que sur les applications avantageuses possibles. Elle présente autant de bons que de mauvais côtés, mais elle a au moins l'avantage de remettre à plus tard, dans le processus d'élaboration, les compromis dictés par la nécessité d'obtenir de meilleures performances et de réduire par le fait même le nombre de ces compromis.

La mémoire auxiliaire a aussi besoin d'une grande capacité parce qu'elle doit emmagasiner d'importants programmes et d'importantes bibliothèques ainsi que des bases de connaissances.

Un poste de travail de l'IA offre habituellement un écran en mode point pour simplifier l'élaboration des interfaces graphiques de haut niveau. Étant donné que les applications qui font appel à l'IA requièrent souvent des interactions complexes avec l'utilisateur du système, il est préférable de pouvoir compter sur le plus grand nom-

bre de techniques d'interface possible afin de simplifier l'utilisation du système.

Le Poste de Travail de l'IA et le TDS

Voici les principales conséquences qui découlent de l'utilisation de cette technologie:

1. Une plus grande productivité du programmeur grâce au meilleur délai d'exécution et aux meilleurs outils offerts par un environnement à haute performances.

2. Une plus grande productivité de l'utilisateur grâce à l'emploi d'interfaces plus naturelles appuyées par une meilleure performance.

3. La possibilité de faire facilement et en plus grand nombre des analyses de données grâce à une plus grande accessibilité à un traitement bon marché. Il pourrait en résulter une meilleure compréhension de nos données et une amélioration des statistiques publiées par la suite.

4. Un traitement meilleur marché rend possible une approche expérimentale permettant une réexécution à l'aide d'autres méthodologies, ou une reproduction pour évaluer l'importance des erreurs introduites par les méthodologies utilisées par des systèmes généraux.

Systèmes Experts

Les premiers travaux de recherche en IA portaient surtout sur des méthodes et des stratégies générales qui se sont avérées incapables de résoudre de vrais problèmes puisqu'elles étaient tout simplement trop coûteuses. Des succès ont été enregistrés lorsque les chercheurs se sont concentrés sur des domaines précis. Le succès remporté par le système de diagnostic des infections bactériennes MYCIN, en particulier, a servi de modèle:

> Créer un système utilisable par des non-informaticiens familiers avec les domaines d'intérêt, et qui incorpore les connaissances glanées par les experts qui ont travaillé dans ce domaine. Le système se sert de ces connaissances pour se comporter comme un expert en dialoguant avec les utilisateurs, en demandant des renseignements et en utilisant les réponses obtenues, ainsi qu'en fournissant spontanément des renseignements et en proposant ses connaissances pour tirer les conclusions appropriées. L'ensemble du procédé est interactif, et les interfaces de haut niveau qui comportent des graphiques techniques et des dialogues en langage naturel contribuent à simplifier l'utilisation du système.

Le système fournit des outils qui permettent aux "experts" de mettre facilement à jour la base de connaissances. De plus, les connaissances sont exprimées sous forme d'instructions que peuvent comprendre les personnes qui connaissent bien le domaine. La base de connaissances et ses fonctions de soutien constituent habituellement une composante distincte et identifiable du système.

Étant donné que les connaissances sont exprimées dans des termes propres au domaine et qu'elles sont vraisemblablement "superficielles" (au sens décrit plus haut), elles ne sont pas immédiatement exécutables. Pour cette raison, le système doit disposer d'une composante qui peut "interpréter" de façon organisée les connaissances exprimées sous forme d'instructions. Cette composante s'appelle généralement le "moteur d'inférence".

Systèmes Experts et TDS

Il paraît peu probable que les systèmes informatiques remplaceront à court ou à moyen terme les "experts" parce que les vrais experts dans n'importe quel domaine, aussi spécialisé soit-il, semblent devoir exercer un certain degré de bon sens général et que la véritable intelligence artificielle n'est vraisemblablement pas pour demain. De plus, les systèmes informatiques (et même les systèmes experts) doivent être mis à jour, et cela suppose que les experts, qui peuvent mettre à jour la base de données, puissent y accéder facilement.

Par contre, les systèmes experts sont très probablement en mesure d'appuyer différents domaines de spécialisation au moyen de programmes de soutien intelligents. Ce sont des programmes machine qui se comportent comme des experts, mais qui ont besoin d'un peu de connais-

sances de base propres au domaine pour pouvoir être avantageusement utilisés. Il en résultera dans l'ensemble une meilleure accessibilité aux compétences de chaque expert. Les experts pourront alors se spécialiser encore plus et leurs compétences, jouir d'une plus grande diffusion. La tâche des groupes d'experts au chapitre de la mise à jour, du partage et du développement de leurs compétences sera également faciliter grâce à l'utilisation d'un système expert.

On pourrait envisager d'appliquer les programmes experts de soutien suivants au TDS:

- Analyste de données de soutien: Ce programme pourrait réunir les compétences d'un analyste de données dans un domaine particulier identifié selon la spécialité et la source. Bien qu'un système expert-analyste de données polyvalent représente probablement un projet trop ambitieux, un programme de soutien possédant des connaissances spécialisées dans un domaine précis, connaitrait les anomalies qui ont déjà causé des problèmes et qui, par conséquent, doivent être filtrées. Il pourrait recommander des techniques d'estimation de variances qui s'appliquent au domaine en question. À titre de complément d'un système général, il pourrait analyser les performances du système, ou suggérer des valeurs de paramètre en se fondant sur l'expérience.

- Économiste de soutien: Ce programme pourrait renseigner un économiste sur les qualités relatives des différentes sources de données requises pour une application précise ainsi que suggérer des techniques pour combiner ou analyser les données. Encore une fois, ce programme, pour être efficace, devrait se concentrer sur un domaine précis.

- Méthodologiste de soutien: Ce programme pourrait être un dépôt de renseignements sur les différentes conceptions d'enquête déjà utilisées, les bases d'enquête disponibles et les systèmes généraux appropriées, qui servirait dans les phases initiales de la conception d'une enquête. Il pourrait aussi donner des renseignements sur les différents aspects de l'analyse des données avant et après l'enquête. La base et le domaine spécialisé, par exemple, devraient être bien définis pour que le programme ne soit pas limité à faire de vagues recommandations.

- Statisticien spécialisé de soutien: Le système expert du statisticien spécialisé pourrait renseigner celui-ci sur des questions comme le contrôle d'une enquête et les possibilités relatives de différentes méthodes de collecte et de saisie des données. Ce programme porterait sans aucun doute sur un domaine spécialisé précis.

- Soutien aux employés de bureau: Beaucoup de travaux traités manuellement sont des travaux courants qui peuvent être exécutés rapidement par une machine. Ce programme éliminerait par filtrage les cas faciles, laissant plus de temps pour les cas plus complexes. Dans l'ensemble, les données traitées seraient probablement plus uniformes, et l'ennui qui caractérise les travaux courants serait réduit.

- Soutien à l'utilisateur: Un utilisateur spécialisé des produits de Statistique Canada sait quelles publications renferment quelles informations et quelle division appeler pour demander des données particulières. Ce programme de soutien spécialisé saisirait ces informations et les mettrait du même coup à la disposition des utilisateurs qui ne sont pas aussi spécialisés.

Une autre sorte de "système expert" qui pourrait devenir très utile est désigné parfois sous le nom de "système frontal à base de connaissances" (SFBC). Ce type de logiciel précise le mode de fonctionnement d'un système informatique, indique si le système convient à la tâche à accomplir, et donne des renseignements sur la façon d'établir différents paramètres pour obtenir une bonne performance. On préconise son utilisation avec d'importantes bibliothèques de logiciels lorsque le choix du sous-programme à utiliser doit passer par un processus décisionnel compliqué. Un SFBD qui se prêterait bien au TDS pourrait être un système général particulier qui exécute une fonction bien définie, et qui est au service d'un groupe d'utilisateurs satisfaits ayant acquis une expérience pratique du fonctionnement du système, y compris de l'établissement de différents paramètres. De plus, s'il y a plusieurs utilisateurs potentiels du système qui sont désavantagés parce qu'ils n'ont pas directement accès à l'ensemble des expériences accumulées, mais qui désirent se servir du système, un système expert pourrait très vraisemblablement constitué un véhicule d'échange de ces expériences.

Un système frontal à base de connaissances, par exemple, pourrait servir à la liaison des articles. Cette opération a souvent posé un problème à Statistique Canada et a entraîné une accumulation des expériences reliées à la théorie et à l'application de celles-ci, y compris la création d'un système général très sollicité (GIRLS). Les méthodologistes se sont prêtés à une approche de type IA qui s'appuie sur un raisonnement approximatif/superficiel, tout en profitant d'une approche expérimentale pour l'analyse des données. De plus, l'application de la technologie à de nouveaux domaines d'activité soulève un intérêt constant qui entraîne une augmentation de la demande pour les compétences existantes. Un système de vérification et d'imputation ainsi qu'un système qui donne des renseignements sur les cas de violation de la confidentialité, représentent d'autres exemples d'application.

La liste ne se veut pas exhaustive. L'important est de savoir où se trouvent les compétences au sein de l'organisation, dans quels cas le domaine est suffisamment spécialisé et quand il peut être avantageux de saisir les connaissances et de les rendre plus accessibles.

Coquilles des Systèmes Experts

Les personnes qui ont créé et mis en application les systèmes experts ont constaté par la suite que si elles enlevaient le contenu de la base de connaissances, elles se retrouvaient avec une portion de logiciel qui pourrait servir à créer un autre système expert. Cette découverte a conduit à la création du système EMYCIN qui correspondait au système MYCIN sans les connaissances sur les infections bactériennes.

Par conséquent, une coquille de système expert devrait offrir un logiciel qui assure la maintenance de la base de connaissances, le moteur d'inférence qui permet d'utiliser les connaissances et une interface de haut niveau. Naturellement, les créateurs de logiciel des coquilles de systèmes experts considèrent maintenant leur produit en termes de fonctionalité, et plusieurs coquilles de systèmes experts qui ne tirent pas leur origine d'un système expert particulier, ont été mis sur le marché. Ces systèmes proposent une grande diversité de fonctions qui peuvent servir à l'élaboration de systèmes experts, mais qui deviennent simplement un environnement de programmation pour un langage très évolué. Les connaissances sont représentées dans un langage déclaratif très évolué à l'aide d'outils de vérification propres au langage (SSBD-connaissances). Le moteur d'inférence interprète le langage puisqu'il utilise les connaissances pour contrôler un processus de calcul qui dans l'analyse finale doit correspondre à un calcul algorithmique. Les outils qui servent à l'élaboration de l'interface sont semblables à ceux que tout bon environnement de programmation devrait offrir comme outils généraux de soutien de bibliothèques.

Les observations qui précèdent suggèrent que les coquilles des systèmes experts doivent être évaluées en fonction de l'environnement de programmation qu'elles offrent ainsi que de la tâche à accomplir:

1. Le langage sous-entendu par le système de représentation des connaissances et le moteur d'inférence, doit être évalué comme d'autres langages, notamment, pour la précision de la sémantique, l'efficacité, etc... Il sera naturellement très évolué pour s'adapter aux applications voulues.

2. L'interface d'édition doit être évaluée comme le sont d'autres éditeurs propres à un langage.

3. Les outils d'interface doivent être de haut niveau, faciles à utiliser et, si possible, transférables; si ce n'est pas le cas, les outils doivent assurer un accès complet aux fonctions de la machine et établir une démarcation claire entre cette caractéristique et la transférabilité.

4. Les coquilles des systèmes experts doivent fournir des outils d'élaboration de logiciel comme des programmes de mise au point, des progiciels de traçage et des outils de mesure des performances. Ces outils doivent ê tre faciles à utiliser et puissants.

CSE et TDS

Les coquilles de systèmes experts peuvent contribuer à améliorer la productivité des programmeurs, surtout de ceux qui élaborent des systèmes experts. Malheureusement, un grand nombre de coquilles de systèmes présentent des problèmes. Ce fut le cas d'importants systèmes truffés d'erreurs et pour lesquels beaucoup de fonctions ne pouvaient pas être mises en oeuvre, rendant difficile le travail des créateurs d'applications. Les

coquilles de systèmes experts n'étaient traditionnellement disponibles qu'avec un matériel spécialisé et ont produit des applications autonomes qui ne peuvent pas être rattachées à d'autres systèmes logiciels.

Ces problèmes ont été solutionnés en partie, mais parce qu'aucune norme n'a été établie, le développement des coquilles de systèmes experts a pris toutes sortes de direction obligeant ainsi l'utilisateur à s'en remettre à un seul fournisseur de logiciels. De plus, le logiciel coûte cher et est difficile à évaluer correctement sans engager des dépenses additionnelles pour installer un prototype suffisamment complexe, et comparer le système livré à d'autres coquilles pour pouvoir apprécier sa maniabilité, son efficacité, sa transférabilité et ses possibilités d'extension.

Par contre, un grand nombre de systèmes à base de connaissances ont été élaborés à partir de coquilles de systèmes experts et le marché pousse les fournisseurs à être plus concurrentiels. Cela entrainera sans aucun doute la création de meilleurs systèmes qui seront mis au banc d'essai et possiblement standardisés, et que l'on pourra se procurer à un coût plus raisonnable.

Conclusions

Les ressources potentielles des technologies de l'IA peuvent être mis à profit pour traiter les données statistiques, ce qui aura pour effet de permettre de créer plus rapidement et à un moindre coût des systèmes généraux plus performants, plus faciles à utiliser et plus adaptables.

Appendice A

EXEMPLES DE PROGRAMMATION À UN TRÈS HAUT NIVEAU

Afin d'avoir un aperçu des trois types de méthodes de programmation, examinons le problème (artificiel) du tri d'une liste.

Dans la méthode de la *programmation fonctionnelle*, la récurrence remplace l'itération, et l'attribution destructive de valeurs aux variables est supprimée. Voici un exemple de mise en application en ML du tri par insertion pour les nombres entiers [14]:

```
fun sort nil = nil
  | sort(a::l) = insert (a,sort l);
fun insert (x:int,nil) = [x]
  | insert(x,l' as a::l)
        = if x <= a
          then x::l'
          else a::insert(x,l);
```

En d'autres mots, le résultat du tri d'une liste vide (`nil`) est une liste vide (`nil`). Le résultat du tri d'une liste dont le premier élément est `a` et le reste `l` correspond au résultat de l'insertion de `a` dans le résultat du tri de `l`. L'insertion de `x` dans la liste vide produit la liste `[x]`. Le résultat de l'insertion de `x` dans une liste `l'` dont le premier élément est `a` et le reste, `l`, dépend de l'équation `x <= a`. Si l'équation est exacte, le résultat est constitué en établissant une liste ayant `x` comme premier élément et `l'` comme reste. Si ce n'est pas le cas, on obtient le résultat en insérant `x` dans la liste plus courte `l` et en constituant un résultat dont le premier élément est `a` et le reste, le résultat de l'insertion de `x` dans `l`.

Des fonctions d'ordre plus élevées permettent d'enlever les dépendances dans la relation de classement intégrée.

```
fun sortgen cfn nil = nil
  | sortgen cfn (a::l)
        = insertgen cfn (a,sortgen cfn
l);
fun insertgen cfn (x,nil) = [x]
  | insertgen cfn (x,l' as a::l)
        = if cfn(x,a)
          then x::l'
          else a::insertgen cfn (x,l);
```

Dans ce cas, `sortgen` est une fonction qui prend un seul argument fonctionnel et retourne un résultat fonctionnel. Si l'argument correspond à une fonction de classement pour les nombres entiers, et que le résultat est une fonction de tri pour les nombres entiers qui se comporte exactement comme dans l'application précédente:

```
fun lesseq(x:int,y) = x <= y;
fun sort list = sortgen lesseq list;
```

Dans la méthode de *programmation logique*, on observe qu'une liste LS d'éléments représente une version triée de la liste L si et seulement si la

liste LS peut être obtenue à partir de la liste L suite à une transformation de permutation (un reclassement), et que la liste LS est dans un ordre de tri.

```
sort(L,LS) :-
    permutation(L,LS),
    ordered(LS).
```

Veuillez noter que cette instruction représente une spécification de haut niveau de la signification du tri et ne présuppose aucun algorithme particulier. Il s'agit également d'une instruction Prolog valable dont la représentation déclarative correspond essentiellement à ce qui est présenté. Les expressions `sort( )`, `permutation ( )`, et `ordered ( )`, sont appelées des prédicats. Un prédicat peut être considéré comme un essai de toutes les paires possibles (L,LS) qui décide si, pour chaque paire, la liste LS est une permutation triée de la liste L.

Le reste du programme de tri se lit comme suit:

```
permutation(Xs,[Z|Zs]) :-
    select(Z,Xs,Ys),
    permutation(Ys,Zs).
```

En d'autre mots, `[Z|Zs]`, une liste ayant `Z` comme premier élément et `Zs` comme reste, correspond à une permutation de `Xs` si `Z` se trouve dans la liste `Xs` et s'il laisse, une fois enlevé, `Ys`, et si `Zs` est une permutation de `Ys`.

```
permutation([],[]).
```

Le cas de base: la liste vide (`[]`) est une (la seule) autopermutation.

```
select(X,[X|Xs],Xs).
```

En d'autre mots, `X` se trouve dans une liste présentant la forme `[X|Xs]` et laisse le reste `Xs`.

```
select(X,[Y|Ys],[Y|Ys]) :-
    select(X,Ys,Zs).
```

Sinon, `X` se trouve dans une liste `[Y|Ys]` laissant `[Y|Zs]` si `X` se trouve dans la liste `Ys` laissant `Zs`.

```
ordered([X,Y|Ys]) :-
    less_eq(X,Y),
    ordered([Y|Ys]).
```

Une liste `[X,Y|Ys]` (dont le premier élément est `X`, le second élément, `Y` et le reste, `Ys`) est dans un ordre de tri si la valeur de `X` est inférieure ou égale à `Y` et que la liste `[Y|Ys]` est triée.

```
ordered([X]).
ordered([]).
```

Deux cas de base: les listes ayant les éléments un et zéro, sont triés.

Le programme qui vient d'être donné est un programme Prolog exact dans la mesure où la valeur de `less_eq/2` (`less_eq` avec deux arguments) a été définie. Toutefois, comment sera-t-il exécuté?

En partant d'un prédicat à prouver (une demande de calcul):

```
sort([2,4,1,3],Answer).
```

Prolog calculera les permutations une à une `[2,4,1,3]`:

```
[2,4,1,3]
[2,4,3,1]
[2,1,4,3]
[2,1,3,4]
[2,3,4,1]
[2,3,1,4]
[4,2,1,3]
[4,2,3,1]
[4,1,3,2]
 ...
```

Chacune de ces permutations sera comparée à la relation "triée" pour déterminer si la liste est triée ou non. L'opération prendra fin lorsque le programme trouvera la permutation `[1,2,3,4]`.

Cet exemple indique une force et une faiblesse du Prolog; elle équivaut à une spécification tout en étant inefficace. Par contre, Prolog peut imiter le comportement d'un meilleur algorithme de tri comme le démontre la mise en application suivante d'un tri par insertion:

```
sort([],[]).
sort([A|L],Result) :-
    sort(L,LS),
    insert(A,LS,Result).

insert(X,[],[X]).
insert(X,[A|L],[X,A|L]) :-
    less_eq(X,A).
insert(X,[A|L],[A|Result]) :-
    not less_eq(X,A),
    insert(X,L,Result).
```

A titre d'exemple de la *méthode orientée objets* nous examinerons le Smalltalk-80. Le Smalltalk-80 comprend une classe d'objets qui sont maintenus dans un ordre de tri appelé "SortedCollection". Une instance de l'objet est créée en envoyant un message `sortBlock:` à la classe

`SortedCollection`, ainsi qu'un bloc, l'argument, qui donne la relation de classement. Cette opération produit un objet vide qui est prêt à recevoir des valeurs. Au fur et à mesure où ils sont envoyés à un objet de cette classe, les éléments sont insérés dans la bonne place.

```
ls <- SortedCollection
    sortBlock: [:a :b | a<=b].
(((ls add: 2) add: 4) add: 1) add: 3
```

Le code même du tri par insertion correspondrait à une méthode rattachée à la classe `SortedCollection`. Pour pouvoir prendre en charge le tri d'une autre classe, l'usage veut qu'une méthode `asSortedCollection` soit fournie pour constituer et retourner un ordre de tri (SortedCollection) à l'aide de ce mécanisme. Par conséquent, l'ensemble formé par les éléments 2,4,1,3 peut être trié en se servant de la commande suivante:

```
#(2 4 1 3) asSortedCollection:
    [:a :b | a <= b]
```

En résumé, la programmation fonctionnelle fournit un langage orienté vers le traitement (sans effets secondaires) qui est facile à utiliser tout en étant efficace. La programmation logique permet au programmeur d'écrire les spécifications ou les descriptions du traitement et de les exécuter. La programmation orientée objets donne une description orientée vers le traitement de l'opération effectuée sur l'objet.

Références

(1) Apple Computer, Inc. (1987). *Apple Human Interface Guidelines: The Apple Desktop Interface.* Addison-Wesley, Reading, Mass.

(2) Fellegi, I. P. and Sunter, A. B. (1969). *A Theory for Record Linkage.* Journal of the American Statistical Association, Vol 64, pp. 1183-1210.

(3) Genesereth, M. R. and Nilsson, N. J. (1987). *Logical Foundations of Artificial Intelligence.* Morgan-Kaufmann, Los Altos, CA.

(4) Goldberg, A. and Robson, D. (1983). *Smalltalk-80: The Language and its Implementation.* Addison-Wesley, Reading, Mass.

(5) Goldberg, A. (1984). *Smalltalk-80: The Interactive Programming Environment.* Addison-Wesley, Reading, Mass.

(6) Harmon, P. and King, D. (1985). *Expert Systems.* Wiley, New York.

(7) Lloyd, J. W. (1987). *Foundations of Logic Programming. (2nd ed.)* Springer.

(8) Parsaye, K. and Chignell, M. (1988). *Expert Systems for Experts.* Wiley, New York, N.Y.

(9) Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, NJ.

(10) Statistique Canada (1987). *Recensement du Canada de 1986: Caractéristiques de la population et des logements: Âge, sexe et état matrimonial.* Statistique Canada, Publication No. 93-101.

(11) Steele, G. L. (1984). *Common LISP.* Digital Press.

(12) Sterling, L. and Shapiro, E. (1986). *The Art of Prolog.* The MIT Press, Cambridge, Mass.

(13) Taylor, W. A. (1988). *What Every Engineer Should now About Artificial Intelligence.* The MIT Press, Cambridge, Mass.

(14) Wikstrom, A. (1987). *Functional Programming using ML.* Prentice-Hall.

UN PRÉVISION TECHNOLOGIQUE DU RÔLE DES ORDINATEURS PERSONNELS AU NIVEAU DES MÉTHODES STATISTIQUES D'ENQUÊTES

J. C. NASH,

L'université d'Ottawa

Résumé

Les prévisions technologiques, c'est-à-dire les prédictions touchant les méthodes et les outils qui seront utilisés dans l'avenir, sont notoirement reconnues pour leur manque de précision. Néanmoins, elles sont souvent utiles aux planificateurs et aux gestionnaires parce qu'elles indiquent la direction éventuelle du développement ou parce qu'elles laissent entrevoir des possibilités ou des dangers susceptibles de se présenter. L'auteur essaie de présenter une prévision technologique pour les ordinateurs personnels et les postes de travail, en ce qui a trait à leur utilisation pour effectuer et analyser des enquêtes statistiques. En particulier, les sujets qui seront abordés pour les ordinateurs personnels seront notamment les possibilités de communication, les logiciels de manipulation des données, les techniques de stockage et la formation des utilisateurs. Ces sujets seront examinés en fonction des tâches relatives à la planification et à la réalisation d'enquêtes statistiques ainsi qu'à l'analyse et à la présentation des données d'enquêtes statistiques.

Introduction

Les prévisions technologiques sont les prédictions des méthodes, des instruments et, de façon plus générale, de tout ce qui s'y rattache. Ces prévisions sont reconnues pour leur manque de précision et pour leurs erreurs flagrantes, tant par intention que par omission, présents dans la publication. Je vais néanmoins essayer de présenter ici quelques prévisions technologiques se rapportant aux postes de travail et aux autres technologies de pointe qui pourraient surgir dans l'exécution des enquêtes statistiques.

Il est important de relever dès le départ les principales lacunes des prévisions technologiques. D'abord, elles sont toujours fausses à un point de vue ou à un autre. Tant la date que la nature des changements sont difficiles à évaluer, comme on le sait bien. Les développements imprévus dans la technologie, la commercialisation, les changements sociaux ou l'action gouvernementale peuvent bouleverser le contexte des progrès technologiques.

Quoi qu'il en soit, les prévisions aident souvent les planificateurs et les gestionnaires en mettant en évidence la direction du mouvement général, ou en montrant des possibilités ou des dangers probables. Dans le secteur de l'agriculture, par exemple, les prévisions de pénurie ou de saturation du marché sont parfois faites et transmises aux producteurs dans le but avoué de permettre de rectifier la production. Les corrections peuvent "fausser" les prévisions, mais on peut dire que ces prévisions se sont révélées utiles.

La présentation qui accompagnera ce document devrait être donnée à la fin du symposium 88. La communication orale mentionnera naturellement les idées présentées par les autres conférenciers qui n'ont pas été retenues ici.

Examinons quelques méthodes de prévision technologique. La plus simple est celle de l'instinct d'une personne qui est reconnue être un expert dans le domaine en question. L'opinion d'un panel d'experts est en général préférée à celle d'une seule personne, et on peut citer comme exemple marquant au Canada la conférence sur les perspectives agricoles. Les panels risquent d'être influencés par une ou plusieurs fortes personnalités qui agissent sur les autres. La méthode Delphi essaie de surmonter ces facteurs psychologiques en isolant les participants. Afin de réviser les opinions, un animateur communique les opinions et les modifications une ou plusieurs fois aux membres du panel et fait rapport sur le consensus ou l'absence de ce dernier à la fin.

Ces méthodes sont essentiellement qualitatives par définition. On peut y ajouter la rédaction de scénarios, dans laquelle une personne ou une équipe essaie de trouver des

directions possibles du déroulement des événements. Il peut s'agir d'un travail purement qualitatif, mais certains travailleurs utilisent des dés ou des choix pseudo-aléatoires générés par ordinateur pour trouver la direction de l'évolution du scénario (Sutherland, 1975). Quoi qu'il en soit, le succès de ces techniques dépend dans une mesure considérable de l'attitude de ces utilisateurs.

J'appelerai "extrapolations" une famille plus quantitative de méthodes. À mesure que la technologie progresse, nous constatons que les diverses mesures de la performance ou des coûts changent avec le temps. Il est possible de procéder à une extrapolation en traçant une courbe sur un graphique de la mesure par rapport au temps. On peut citer comme exemples de graphiques de ce genre et de leurs extrapolations le coût par kilobit-mémoire, les circuits de transistor par puce ou le coût par méga-octet d'espace de mémoire sur disque. Le principal inconvénient de ces extrapolations est qu'elles peuvent laisser de côté l'introduction d'une forme de technologie tout à fait nouvelle ou encore ignorer les limites fondamentales du progrès technique, bien que les utilisateurs informés tiennent compte généralement de cette dernière possibilité.

Les courbes de croissance sont un cas particulier des extrapolations, et sont particulièrement utiles pour la prédiction de l'adoption de nouvelles techniques. En particulier, on peut utiliser la courbe sigmoide (logistique, de Gompertz, ou autre) pour modeliser la proportion d'une population qui a opté pour une nouvelle technologie. J'ai une expérience personnelle de cette approche dans le domaine de la production laitière (Nash and Teeter, 1975; Nash, 1977, Nash and Walker-Smith, 1987). Des techniques semblables existent pour un grand nombre d'autres applications (par exemple Fisher and Fry, 1983). L'inconvénient de cette méthode est que nous avons en général davantage besoin de plus de données que celles qui existent afin d'obtenir la prédiction de la position de l'asymptote et de la date de réalisation à une certaine distance de cette dernière.

Afin de régler le problème de l'absence de données, les prévisionnistes peuvent substituer dans le modèle de croissance des paramètres tirés de modèles de changements technologiques semblables qui se trouvent plus loin dans le processus de changement. L'inconvénient ici est le mot "semblables", avec l'hypothèse peut-être risquée que le taux d'adoption de "nouvelles" idées suivra la même tendance dans deux situations différentes.

Une idée semblable est d'utiliser une mesure (variable) comme prédicteur d'une autre, comme le concept des indicateurs dominants (Makridakis et al., 1983, chapitre 12). Même si les analystes boursiers aiment beaucoup ces méthodes, je pense que les mesures choisies comme indicateurs sont soit évidentes, ce qui veut dire qu'elles fournissent très peu d'informations quant à l'avenir que l'on ne connaisse déjà, soit qu'elles sont valables pour un nombre trop limité d'événements ou d'intervalles de temps pour être vraiment utiles sur le plan pratique.

Comme on peut le constater, toutes les méthodes quantitatives mentionnées jusqu'à présent sont applicables principalement à court terme. Pour le plus long terme, c'est-à-dire plus de trois à cinq ans, il y a peu de méthodes. Les modèles structurels à long terme, principalement les modèles économétriques, offrent une possibilité. Cependant, ils tiennent rarement compte de l'évolution technologique. De fait, ils retiennent l'hypothèse plus ou moins explicite que les procédés et les tendances du passé se reproduiront dans l'avenir.

Une méthode qui peut permettre au prévisionniste d'obtenir quelques renseignements quantitatifs sur l'évolution technologique à long terme est la modélisation à incidence recoupée (Turoff, 1975, et autres références ci-dessous). Ceci peut être à la fois un instrument conceptuel et de calcul. L'idée fondamentale est que la mise au point d'une technologie améliore ou diminue la vraisemblance ou le taux de mise au point d'une autre.

On peut citer comme exemple les caisses enregistreuses électroniques, qui ont mis un terme aux appareils électro-mécaniques.

Les principales lacunes de ce type d'analyse sont que l'utilisateur peut ne pas inclure une technologie valable ou peut fournir de mauvaises estimations de la pondération correspondante. Cependant, ce type d'analyse améliore les conjectures quant à l'avenir, et sera la principale forme d'analyse utilisée ici. Cependant, j'estime que cette approche, sous sa forme actuelle, et telle qu'elle est utilisée ici tout au moins, est avant tout un instrument de structuration de mes pensées. En préparant ce document j'ai essayé de mettre au point un programme pour détailler et élargir les idées qui seront examinées, mais j'estime qu'il y a tellement de choix et d'hypothèses dans la mise en oeuvre de ce programme que des considérations d'espace et de temps empêchent de les mentionner davantage ici.

Cette présentation va essayer de tirer des prévisions technologiques pour l'impact des ordinateurs personnels et des postes de travail dans leur application éventuelle à l'exécution et à l'analyse d'enquêtes statistiques. L'horizon temporel de ces prévisions est l'an 2000, c'est-à-dire dans 12 ans.

Enquêtes

Le tableau 1 présente une vue d'ensemble des technologies de postes de travail propres aux différents aspects d'une enquête. Les types de logiciels qui pourraient être utilisés sont également énumérés. Je crois que des exemples de tous les types de logiciels mentionnés existent, mais je pense qu'aucun système consolidé complet n'a encore été mis en oeuvre, même si le fondement technologique est là. Ceci peut s'expliquer en partie par des conflits ou des incompatibilités des structures de données, des normes et des milieux informatiques utilisés dans différents sous-systèmes. Il est plus probable que les ressources organisationnelles, de traitement des données et de personnel existantes, ainsi que les restrictions budgétaires, empêchent la mise au point d'un système d'enquêtes statistiques à poste de travail unifié.

Dans le cas des enquêtes qui ne portent pas sur les questions des répondants, nous pouvons observer des changements très intéressants de la méthodologie d'enquête. Examinons par exemple l'interrogation d'instruments par télécommunication. On peut le faire, et on l'a déjà fait, grâce au modem habituel avec capacité d'auto-réponse. On envisage déjà la possibilité de lire à distance un compteur de services publics. Normalement, de telles applications seraient régulièrement utilisées à des fins de facturation, mais on peut envisager leur utilisation dans une enquête par sondage afin de déterminer la nature de la demande de services publics une fois que le matériel approprié sera en place.

Un tel exemple fait ressortir plusieurs des questions qui seront soulevées plus loin:

1) les utilisateurs peuvent considérer que l'utilisation d'un tel matériel dans l'enquête peut être une intrusion dans leur vie privée

2) le coût de l'investissement est tel que des utilisateurs pourraient ne pas recevoir le nouveau matériel

3) les employés sont remplacés par du matériel et viennent grossir les rangs des chômeurs.

En dépit du risque d'utilisations malencontreuses d'enquêtes automatisées de ce genre, j'estime qu'au cours des 15 prochaines années le nombre de leurs applications va croître. On peut citer comme domaines particulièrement visés la surveillance de l'environnement

(intérieur et extérieur), le contrôle de la qualité et l'amélioration de la fabrication des services, le mouvement des humains et des véhicules, la gestion de la faune ou du bétail, la vérification financière et des enquêtes semblables de la comptabilité informatisée ou des dossiers de points de vente, ainsi que les enquêtes sur la commercialisation et la publicité à partir des programmes de télévision.

Flux de technologie existants

L'extrapolation des flux existant dans la mise au point des postes de travail et des techniques connexes est relativement directe. On peut citer comme exemple de cette extrapolation le Tablet, qui est un ordinateur portatif (Mel et al., juin 1988). C'est un produit pour l'an 2000. Cependant, sur la couverture du numéro du 1er août 1988 de Fortune Magazine, on peut voir John Scully, président de Apple Computer, tenir un prototype du Knowledge Navigator, dont l'aspect et la description font qu'il semble être une implémentation de Tablet.

L'une des importantes caractéristiques de Tablet est l'utilisation éventuelle d'un écran, sur lequel l'utilisateur écrit avec un stylet. Le logiciel de reconnaissance de l'écriture manuelle est conçu pour "apprendre" l'écriture de l'utilisateur et le traduit en temps réel. L'utilisateur est donc en mesure de corriger des erreurs de traduction, ce qui réduit la nécessité pour le logiciel d'avoir un taux d'erreur extrêmement bas. Tablet a des applications évidentes à l'étape de l'interview, mais offre de nombreuses autres possibilités.

Des extrapolations plus conventionnelles touchent les domaines des communications, de la manipulation des données, de la technologie de stockage et de la formation des utilisateurs.

Les installations de communication pour la transmission des données s'améliorent de façon soutenue. En dépit des problèmes posés par la sécurité et la confidentialité, il semble probable que les enquêtes statistiques vont utiliser de plus en plus la communication de données plutôt que la voix pour la collecte des données. D'autres possibilités importantes s'ouvrent au travail en équipe dans l'analyse des données d'enquête par des travailleurs à des endroits différents et à la diffusion des résultats de l'enquête aux clients.

La sophistication accrue du logiciel DEVRAIT améliorer la manipulation des données. Cependant, les intérêts commerciaux pourraient empêcher le mouvement facile des données. On peut citer comme exemple l'impossibilité pour certaines versions de Lotus 1-2-3 de lire les fichiers de feuilles de travail créés par d'autres versions du même progiciel. La raison de cette incompatibilité semble être la vente de versions "améliorées" du logiciel, mais cette question se trouve estompée par la présence d'améliorations pour les versions plus nouvelles ou non didactiques. D'autres progiciels peuvent avoir des conflits de versions semblables.

La capacité par dollar du dispositif de stockage de données a enregistré une croissance soutenue. On peut facilement le constater dans le cas des dispositifs à disque magnétiques souples et fixes. Le stockage des données optiques était annoncée depuis un certain nombre d'années, mais sa diffusion commerciale et les fluctuations de prix semblent être beaucoup plus lentes que prévu d'après les premiers documents de presse et les commentaires. Néanmoins, il y a des dispositifs qui emmagasinent 200-1000 méga-octets de données, ce qui permet aux stations de travail d'avoir accès aux déclarations d'enquêtes très importantes et d'emmagasiner des images graphiques qui représentent par exemple, les rues d'une municipalité à enquêter.

L'éducation des utilisateurs est un domaine dans lequel les perspectives des ordinateurs personnels n'ont pas encore été entièrement étudiées, loin de là. Le logiciel d'enseignement assisté par ordinateur (E.A.O.) s'est révélé très cher à préparer et à installer. (À la réunion

annuelle de la Société statistique du Canada de 1988 à Victoria, M. J.C. Holt de l'Université de Guelp, a été l'un des participants qui ont décrit les efforts pour mettre au point un tel logiciel). Il semblerait que l'E.A.O. pour différents aspects des enquêtes par sondage serait un dispositif extrêment efficace. Les participants potentiels d'un tel cours de formation, qui s'attache peut-être à un sous-domaine des enquêtes statistiques, sont souvent très éloignés les uns des autres et/ou ont des horaires qui ne leur permettent pas de se rassembler dans un endroit et à une date uniques pour recevoir la formation en question. En raison du coût de la mise au point des cours E.A.O., des aspects sociaux de l'enseignement et du fait qu'il a peut-être tendance à transformer le concepteur en chômeur, je ne prévois pas de progrès rapides en ce domaine. De fait, je soupçonne qu'en l'an 2000 nous verrons que l'E.A.O. a un rôle plus important, mais toujours très faible, dans le domaine de l'éducation principalement, et il ne prendra sa place au sein de la formation en enquêtes statistiques que lorsque des efforts spéciaux seront consentis en ce sens.

Impact de la technologie sur les enquêtes

Je pense que l'impact direct des progrès dans le domaine des postes de travail et des technologies connexes est assez facile à prévoir. La marche régulière et peut-être inexorable du progrès sur le plan de la vitesse et de la capacité des postes de travail fera naturellement qu'une très petite équipe statistique sera en mesure d'effectuer des enquêtes assez importantes sans interviews personnelles, c'est-à-dire celles qui sont faites par la poste ou par communication électronique, ou qui ne nécessitent aucune main-d'oeuvre. Dans la partie précédente, j'ai déjà présenté une évaluation technologique plus détaillée. Les questions qui restent semblent être celles de l'intégration du personnel, du logiciel, du matériel et des procédures d'exploitation dans des systèmes qui fonctionnent sans heurts. En d'autres termes, je pense que les principaux obstacles sont ceux des incompatibilités des critères et des capacités des enquêtes et des moyens de les exécuter.

Il est moins facile de prédire l'évolution de l'impact plus généralisé de ces considérations. Cependant, avec la témérité d'une personne dont l'existence n'est pas directement tirée des enquêtes, je vais tenter de présenter quelques prévisions, en précisant cependant au départ que je ne suis pas encore en mesure de fournir des preuves quantitatives.

D'abord, il est probable que le coût global de l'exécution d'une enquête, de l'étape de la planification à celle de la publication, devrait diminuer avec la baisse du personnel que la technologie des postes de travail permet. Les pressions exercées au sein des gouvernements pour réduire le nombre de fonctionnaires peuvent trouver un exutoire grâce à cette technologie. Une pression de sens opposé existe cependant dans le cas de la tendance vers le raccordement des programmes gouvernementaux, ce qui comprend les paiements de transfert de divers types, aux résultats des enquêtes. Au Canada, les enquêtes sur les coûts de production servent à fixer les prix de certains produits alimentaires, pour ne citer qu'un exemple de cette tendance. Les applications informelles des résultats d'enquête sont aussi largement utilisées par les représentants élus et les groupes d'intérêts pour réclamer des mesures destinées à corriger les désavantages perçus dans l'emploi, les revenus, l'éducation, les services médicaux, etc. Il est clair que les statistiques sont importantes.

Ensuite, en raison de cette importance, nous pouvons nous attendre à ce que des initiatives soient prises afin que les résultats ou les interprétations des enquêtes viennent confirmer un point de vue particulier. Avec la diffusion de la technologie électronique et des connaissances sur les plans de la construction et du fonctionnement, il semble que l'altération des enquêtes à des fins économiques ou politiques posera des problèmes aux statisticiens. Je ne veux pas dire qu'un tel sabotage sera généralisé, mais je veux prévenir que la concentration du contrôle des enquêtes dans les mains d'un nombre moins grand de personnes et l'accroissement des opérations électroniques faciliteront une telle activité.

Même sans aucune interférence dans l'exécution de l'enquête ou avec les données, la technologie actuelle fait qu'il est plus facile et moins onéreux de "former" les répondants.

Il y a quelques années au Canada, un groupe à intérêt spécial a tenu un séminaire sur la façon de répondre à un questionnaire d'enquête afin de fournir les résultats les plus favorables au groupe. À part ces falsifications pour tirer un avantage particulier, nous pouvons nous attendre à ce que des vandales ou ceux qui sont intéressés à empêcher l'enregistrement de l'information détruisent cette dernière.

En cas de destruction ou d'altération de l'information, je soupçonne qu'une importante partie de la population va en fait applaudir les coupables, au lieu de s'indigner. Les inquiétudes quant au respect de la vie privée et des droits de la personne humaine sont une préoccupation croissante (Dyer, 1988). Le fait que le questionnaire du recensement du Canada ait été rejeté en cour (voir les références ci-dessous), quels que soient ses mérites juridiques ou moraux, sert à illustrer ce point. Le raccordement des fichiers par les organisations gouvernementales ou privées afin de tirer des renseignements sur les citoyens se trouve considérablement favorisé par la mise au point d'un logiciel d'intelligence artificielle et de matériels nouveaux tels que la Connection Machine (Stanfill and Kahle, 1986). De même, la capacité des organismes d'effectuer des enquêtes sans un questionnaire en bonne et due forme, ni même sans que le répondant s'en rende compte, est une possibilité que le grand public n'est pas prêt d'accepter facilement, et il se méfie de la capacité des grands organismes de traiter les données touchant les personnes de façon appropriée.

En dépit des préoccupations touchant les "droits" des individus ou tout au moins, les privilèges, notre société semble se partager en deux parties principales: ceux qui possèdent une propriété et touchent un revenu, et ceux qui vivent en marge, c'est-à-dire les sans-logis, les handicapés, les criminels, les drogués et les débiles mentaux. J'appelle cette société marginale au tableau 2 les démunis, en l'absence d'un terme plus générique. Les progrès technologiques peuvent accroître ce schisme dans notre société en rendant certaines personnes incapables de gagner leur subsistance, à mesure que leurs vrais emplois sont confiés à des machines ou sont dégradés au point qu'ils payent trop peu si on les compare au bien-être social ou au crime. En ce qui concerne le statisticien qui essaie d'effectuer une enquête, les démunis posent un défi formidable. On ne peut facilement les placer dans une base de sondage, ils peuvent réagir violemment aux tentatives d'interview et ils peuvent fournir les réponses qui leur semblent les plus valables personnellement. Les politiques gouvernementales actuelles dans les principaux pays occidentaux ne semblent pas destinées à stopper la tendance à l'augmentation du pourcentage de la population marginale. La bibliographie ci-après contient plusieurs références, dont deux semblent indiquer une enquête auprès des sans-logis, mais au moment de la rédaction de cette communication, je n'ai pas encore été en mesure d'évaluer la méthodologie de cette enquête.

En résumé, les statisticiens doivent s'attendre à ce que la planification, l'analyse et la publication dans les enquêtes soient plus faciles et à ce que l'obtention des données adéquates soit beaucoup plus difficile. Cependant, lorsque des données de qualité sont recueillies, les instruments utilisés devraient permettre l'extraction de meilleurs renseignements d'enquêtes, avec une réduction des délais de publication, une augmentation de la précision et de la fiabilité des estimations et une analyse plus fine en termes de période, de région géographique ou d'autres catégories. De plus, la publication peut, et va, comprendre de meilleurs aides graphiques et devrait éviter des erreurs typographiques par la mise en page directe des tableaux dans le document d'enquête à partir de l'analyse la plus récente.

Conclusion

Au moment où ces prévisions vont paraître dans les actes de cette conférence, je me trouve

dans la situation délicate que mes propos puissent me retomber dessus à une réunion ultérieure. Dans la mesure où une partie de la discussion ci-dessus constitue un avertissement de ce que nous, comme statisticiens, pourrions désirer éviter ou prévenir, j'espère sincèrement qu'une partie de mes prédictions ne se réalisera pas.

Remerciements

La recherche faite par l'auteur est en partie financée par des subventions du Conseil de recherches en sciences naturelles et en génie du Canada. Les communications Telex pour cette contribution ont été fournies par le ministère de la Recherche scientifique et industrielle de Nouvelle-Zélande, Mount Albert Research Centre, Auckland. Les entretiens avec Kevin Price, Gordon Sande, Elaine Hoskins et Mary Nash ont été précieux pour élaborer certaines des idées présentées.

Bibliographie

References

Alberta (Western) Report, "Beating Back Census", v. 14, n. 48, Nov. 16, 1987, p. 42 and v. 15, n. 17, April 11, 1988, p. 24.

Calgary Herald, "Businessman wins fight over census, March 23, 1988, p. B6.

Cetron, M.J. and Ralph, C.A., Industrial Applications of Technological Forecasting: Its Utilization and Management, Malabar, Florida: Robert E. Krieger Publishing Co., 1983.

Dalby, James F., "Practical refinements to the cross-impact matrix technique of Technolgical forecasting", contribution XV in Cetron and Ralph (1983).

Dyer, Gwynne, "Human rights -- and who is human", Ottawa Citizen, Tuesday August 16, 1988, p A8.

Fisher, J.C. and Fry, R.H., "A simple substitution model of Technological change", contribution XVII in Cetron and Ralph (1983).

Fortune Magazine, August 1, 1988 issue, Cover photo.

Halifax Chronicle Herald, "Halifax homeless on increase", Nov. 29, 1986, p. 1,2.

Jantsch, Eric, Technological Forecasting in Perspective, Paris: OECD, 1967.

Linstone, H.A. and Turoff, Murray, The Delphi Method: Techniques and Applications, Reading, MA: Addison-Wesley, 1975.

Makridakis, S., Wheelwright, S.C., and McGee, V.E., Forecasting: Techniques and Applications, 2nd edition, New York: Wiley, 1983

Mel, B.W. et al., "Tablet: personal computer in the year 2000" Comm. ACM, v. 31, n. 6, pp 638-646, June 1988.

Montreal Gazette, "Census violates privacy, commission rules", June 21, 1988, p. A7.

Nash, J.C. "A discrete alternative to the logistic growth function", Applied Statistics, v. 26, n. 1, 1977.

Nash, J.C., and Teeter, N.J., "Building models: an example from the Canadian dairy industry", Canadian Farm economics, v. 10, n. 2, pp. 17-24, April 1975.

Nash, J.C. and Walker-Smith, M., Nonlinear Parameter Estimation: an Integrated System in BASIC, Marcel Dekker Inc.: New York, 1987.

Stanfill, C. and kahle, B., "Parallel free-text search on the Conection Machine system", Communications of the ACM, v. 29, n. 12, December 1986, pp. 1229-1239.

Sutherland, J.W., "Architecting the future: a Delphi-basd paradigm for normative system-building", contribution VI.E. in Linstone and Turoff (1975).

Toronto Globe and Mail, "Reagan's cutbacks blamed as US homeless toll grows", Feb. 6, 1987, p. A10.

Toronto Star, "Group plans survey of homeless", Jan. 14, 1987, p. A3.

Toronto Star, "100,000 Canadians homeless during 1986", April 7, 1987, p. A8.

Toronto Star, "Shelters forced to turn away homeless", April 8, 1988 p. A1, A4.

Toronto Star, "200,000 families said to lack decent housing", June 28, 1988, p.A3.

Turoff, Murray, "An alternative approach to cross impact analysis", contribution V.C. in Linstone and Turoff (1975).

Winnipeg Free Press, "Cutbacks increase (Edmonton's) homeless", August, 10, 1987, p.40.

Winnipeg Free Press, "250,000 Canadians homeless -- national survey", Oct. 23, 1987, p.10.

Winnipeg Free Press, "8 Manitobans in hot water over census", May 3, 1987, p.4.

Tableau 1: Besoins technologiques des postes de travail pour les éléments des enquêtes statistiques.

Éléments des enquêtes:	Instruments informatiques
-planification	
-lancement du projet: besoins, buts	traitement de texte
-planification de l'instrument d'enquête	traitement de texte
-plan de sondage	Calculs, simulation
-mise au point de la base de sondage	base de données, mémorisation
-coûts	chiffrier électronique
-collecte de données	
-"interview" ou équivalent	logiciel spécialisé
-imputation	base de données
-surveillance du déroulement (amélioration de la qualité)	gestion du projet, logiciel spécialisé
-formation du personnel	E.A.O.
-simulations spécialisées	logiciel de simulation + améliorations
-gestion	gestion de projet
-analyse des données	
-agrégation	base de données
-contrôle	base de données
-imputation	base de données (spécialisée)
-détection et affichage des valeurs aberrantes	graphiques/calculs/base de données
-surveillance du déroulement et mise au point du fichier-journal spécialisés ou généraux	instruments de gestion de projet
-traitement de l'estimation	divers
-préparation du document	
-texte	traitement de texte) intégration
-graphiques	instruments graphiques) intégration
-tableaux	chiffrier électronique) intégration
-interprétation	systèmes d'experts

Tableau 2. Technologies ou situations envisagées

1. Améliorations informatiques -- Matériel plus puissant et logiciel avec meilleur interface

2. Démunis -- sans-logis, chômage chronique, analphabétisme, dégradation des emplois, etc.

3. Droits et privilèges des individus -- croissance des préoccupations en matière de respect de la vie privée, de la sécurité personnelle, des libertés individuelles de la discrimination et de l'environnement

4. Critères quantitatifs minimums -- exigence légale ou autre critère pour les résultats des enquêtes statistiques

5. Améliorations des communications -- transfert facilité de l'information, compatibilité des systèmes de données, des normes, etc.

Tableau 3. Impact réciproque des différentes "technologies" les unes sur les autres sur une échelle de -2 à 2 (—, -, 0, +, ++)

L'ampleur de l'effet de la technologie à gauche sur la technologie au haut du tableau est donnée dans l'entrée dans le corps du tableau.

	Améliorations informatiques	Démunis	Droit de la personne humaine	Critère quantitatif	Communications
Améliorations informatiques	++	+	-	++	++
Démunis	0	+	+	--b	-b
Droits de la personne humaine	--	-a	+	-	-
Critère quantitatif	+	0	-	+	+
Communications	+	+	0c	+	+

a -- cet élément est négatif, puisque l'atténuation des inquiétudes à propos des droits et des privilèges de la personne humaine vont RÉDUIRE les cas de dénuement.

b -- l'incapacité de placer les personnes dans une base de sondage fera qu'il sera très difficile de les inclure dans une enquête. Toutefois, on peut penser que ceux qui ne peuvent se permettre tout produit de la société n'ont pas besoin d'être pris en compte. Ici, j'ai posé par hypothèse que tous les citoyens font partie de la société.

c -- la technologie des communications peut avoir des conséquences profondément négatives sur les droits et les privilèges, mais en même temps elle peut être utilisée de façon positive. On a cité à l'auteur plusieurs cas où de graves violations des droits de la personne humaine ont été évitées ou minimisées grâce aux membres d'Amnesty International et par des partisans qui ont utilisé de façon intelligente la technologie des communications.