STATISTICAL USES OF ADMINISTRATIVE DATA

Proceedings

AN INTERNATIONAL SYMPOSIUM

NOVEMBER 23-25, 1987

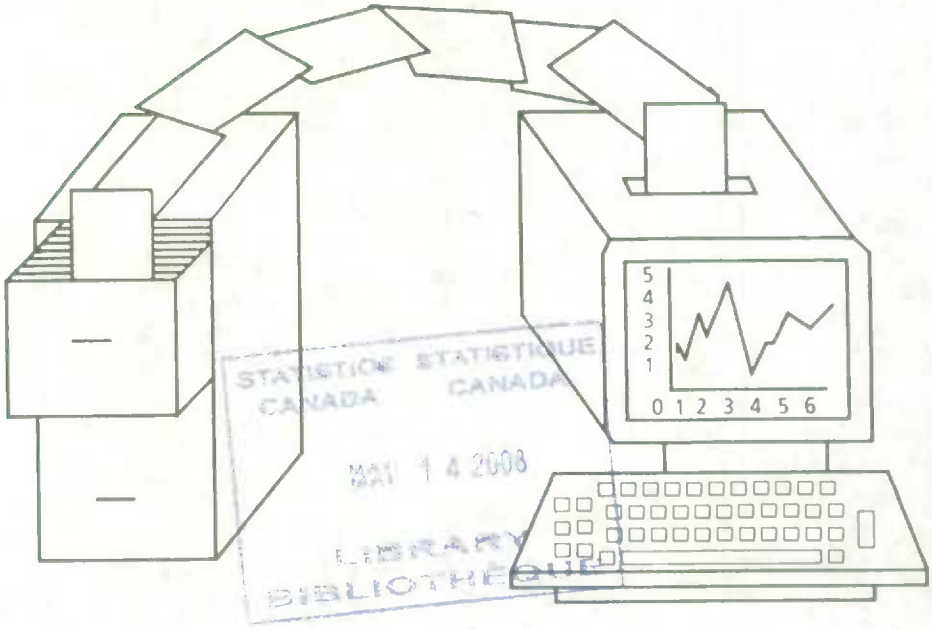
Organized by Statistics Canada

Statistics Canada Statistique Canada

Canada

STATISTICAL USES OF ADMINISTRATIVE DATA AN INTERNATIONAL SYMPOSIUM

EDITED BY
J.W. COOMBS and M.P. SINGH



STATISTICS CANADA
DECEMBER 1988
OTTAWA

PRICE: Canada $35.00
Other Countries: $42.00
Payment to be made in Canadian funds or equivalent

Version française de cette publication est disponible

PREFACE

This symposium on the Statistical Uses of Administrative Data was the fourth in a series of annual symposia on methodological issues organized by Statistics Canada. Topics for earlier symposia were: Analysis of Data from Complex Surveys (1984), Small Area Statistics (1985) and Missing Data in Survey (1986).

Statistical agencies in many countries are facing the challenge of meeting demands for more frequent and detailed data while reducing both response burden and costs. This has led to more use of administrative data in a statistical context. This symposium brought together experts from academic, private and public sectors of Canada, the United States and other countries to share their experience in issues concerning the use of administrative files. Over three hundred people participated in the conference which included six invited paper sessions, a panel discussion, and four contributed paper sessions.

This report organizes edited versions of the papers presented at the conference into ten Sections. Section I presents the organizational experiences and concerns resulting from the increased use of administrative data in various countries. It discusses technical and administrative issues, challenges, confidentiality/privacy concerns, and the potential for register based population censuses. Also reviewed are the recent trends related to the future development of administrative records for statistical purposes.

Considering the Fellegi-Sunter model as the primary model for record linkage, several ways of improving upon this model in applications for cross sectional and longitudinal research, for business registers and for demographic research are presented in the five papers included in Section II. Papers included in section III deal with the use of administrative data at various stages of survey operations such as frame construction, supplementing or replacing survey data, improving estimation methods, data analysis and evaluation and improving data bases by the combined use of data derived from several related surveys and administrative files.

Section IV contains four papers which compare strengths and weakness of different administrative files in producing statistical information, and their use in quality evaluation or the improvement of survey data. Section V presents experiences and studies on the use of administrative data in population censuses, on the impact of changes in administrative data files on the ongoing programs and on the timeliness of such data.

Sections VI to IX contain papers from contributed paper sessions and deal with a wide variety of topics related to the use of administrative files including their use in estimating small area statistics and components of population changes, in balancing for survey nonresponses, in the creation of a dynamic database for students, in determining and updating selection probabilities in the profiling of economic entities, in the production and analysis of crime and medical statistics, and in a quality assurance program.

The last section (Section X) presents the contributions of the four panelists made at the Panel Discussion Session. This session dealt with future challenges in methodological development, protection from invasion of privacy, timeliness and quality of data.

This symposium was successfully organized by a committee consisting of J.W. Coombs and M.P. Singh (Chairpersons), N. Kopustas, F. Mayda, and C. Patrick with the able support of many resource persons from Statistics Canada.

The contents of the papers included in this report are the responsibility of the authors. Papers were not refereed. However, editorial checks were made during the compilation of the papers and any proposals for significant changes were brought to the attention of the authors. Minor editorial changes were considered the prerogative of the editors. The formatting of the papers generally follows the layout and the style used for **Survey Methodology**, a Journal of development and applications of methods in surveys produced by Statistics Canada.

Acknowledgements: The Editors would like to thank Judy Clarke, Josée Dufresne, Dula Edirisinghe, Lucie Gagné, Myra Kent, and Christine Larabie of Methodology Branch of Statistics Canada for their extensive contributions to the copy preparation process. Thanks are also due to the authors and their support staff who made extra efforts in preparing their manuscripts in the required format and for providing us with diskettes.

Special appreciation is due to Christine Larabie and Frank Mayda for their countless hours of effort in making this volume a reality.

Gordon J. Brackstone
Assistant Chief Statistician
Statistics Canada

STATISTICAL USES OF ADMINISTRATIVE DATA

SYMPOSIUM PROCEEDINGS

CONTENTS PAGE

Preface ..... (i)

OPENING REMARKS

I.P. Fellegi, Chief Statistician of Canada 1

INVITED PAPERS

SESSION I: Policy Issues and Organizational Experience
Chairperson: J. Ryten, Statistics Canada

Statistical Uses of Administrative Data: Issues and Challenges, G. J. Brackstone (Statistics Canada) 5

European Experience of Using Administrative Data for Censuses of Population: The Policy Issues that must be Addressed, P. Redfern (United Kingdom) 17

Protection of Taxfiler Data, H.J. Lagassé (Revenue Canada Taxation) 35

Statistical Uses of Administrative Records in the United States: Where are we and where are we going?, F. Scheuren and T. Jabine (U.S. Internal Revenue Service) 43

SESSION II: Record Linkage Methodology
Chairperson: J.N.K. Rao, Carleton University

Using Large Data Bases for Research on Surgery, L.L. Roos and N.P. Roos (University of Manitoba) 75

Missing Identifiers and the accuracy of individual Follow-up, M.E. Fair and P. Lalonde (Statistics Canada) 95

Computational Aspects of Applying of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses, W.E. Winkler (U.S. Bureau of the Census) 109

Concepts and Practices that Improve Probabilistic Linkage, H.B. Newcombe (Consultant), M.E. Fair, and P. Lalonde (Statistics Canada) 127

Record Linkage-Methodology and its Application, M. Eagen (Goss, Gilroy & Associates Ltd), and T. Hill (Statistics Canada) 139

SESSION III: Integrated Approaches to Data Development
Chairperson: B. Petrie, Statistics Canada

Use of Administrative Data in the Business Survey Redesign Project, M. Colledge (Statistics Canada) 153

Small Area Estimation of Employment in Different Classes of Hours Worked, S. Lundstrom (Statistics Sweden) 169

Methodology for Construction of Address Registers Using Several Administrative Sources, J.D. Drew, J. Armstrong, A. van Baaren and Y. Deguire (Statistics Canada) 181

Multiple Uses in Statistics of Administrative Records in the Analysis of Education Data, C.D. Cowan and M.K. Batcher (U.S. Center for Education Statistics) 191

The Social Policy Simulation Database An Example of Survey and Administrative Data Integration, M.Wolfson, S. Gribble, M. Bordt, B. Murphy and G. Rowe (Statistics Canada) 201

SESSION IV: Quality Evaluation
Chairperson: N.P. Gendreau, Bureau de la Statistique du Québec

Data on the Elderly - A Comparison of Two Sources, N.J. Kopustas (Statistics Canada) 233

A Two-Stage Survey: The Permanent Sample of Social Insurance Beneficiaries in France, A. Mizrahi and A. Mizrahi (CREDES, France) 239

Corporation Income Tax Records Used for Tax Policy Analysis, F. Hostetter, C.D. McCann and B. Zirger (Revenue Canada) 249

Using Administrative Record Data to Evaluate the Quality of Survey Estimates J.C. Moore and K.H. Marquis (U.S. Bureau of the Census) 255

SESSION V: Administrative Records as an Alternate Data Source
Chairperson: F. Scheuren, U.S. Internal Revenue Service

Administrative Data as Alternative Sources to Census Data, J. Podoluk (Consultant, Canada) 273

The Quality of Administrative Data from a Statistical Point of View Some Danish Experience and Considerations, P. Jensen (Danmarks Statistik) 291

Evaluating the Effect of Tax Reform on Census Bureau Programs, G. Gates (U.S. Bureau of the Census) 301

CONTRIBUTED PAPERS

SESSION VI: Chairperson: M.P. Singh, Statistics Canada

A Review of the Use of Administrative Records in the Survey of Income and Program Participation, C. Bowie and D. Kasprzyk (U.S. Bureau of the Census) .. 315

Time Series Modelling for Small Area Estimation, G.H. Choudhry and L.A. Hunter (Statistics Canada) 327

Turning the Tables: Imputing for Item Nonresponse When Donors are Scarce, J.L. Czajka (Mathematica Policy Research, Inc. U.S.) 339

SESSION VII: Chairperson: Daniel Kazprzyk, U.S. Bureau of the Census

Relationships of Murder Characteristics to Trial Outcomes and to Capital Punishments, Canada, 1961-1983, J.F. Gentleman and P.B. Reed (Statistics Canada) 355

The Use of Administrative Records in Canada for Estimating Population and Components of Population Change, R.B.P. Verma and R. Raby (Statistics Canada) .. 363

Statistics on Administrative Registers in Mexico Present Situation and Problematic, M.A. Elena Figueroa M. (Statistics Mexico) 373

SESSION VIII: Chairperson: John Coombs, Statistics Canada

Updating Tax Return Selection Probabilities in the Corporate Statistics of Income Program, S. Hinkins, H. Jones and F. Scheuren (U.S. Internal Revenue Service) 379

The Use of Administrative Data for Initial and Subsequent Profiles of Economic Entities, C. Clark and R. Lussier (Statistics Canada) 387

Integrating Student Records into a Dynamic Database and Statistical Reporting System, A.E. Hollings and B.D. Pettigrew (University of Guelph) 401

SESSION IX: Chairperson: Geoff Lee, Australian Bureau of Statistics

Automated Quality Assurance Processing of Administrative Record Files, P. Hanczaryk and J. Jonas (U.S. Bureau of the Census) 413

An Algorithm for the Determination of Optimum Matching Rules for the Linkage or Records from Two Sources, Y. Yesilcay (Sultanate of Oman) 423

SESSION X: Invited Panel Discussion
Chairperson: G. Brackstone, Statistics Canada

Notes for a Panel Discussion on Statistical Uses of Administrative Data, J.W. Grace (Privacy Commissioner of Canada) 433

Remarks at Panel Discussion, T.B. Jabine (Consultant, United States) 437

Privacy Issues Involved in the Exploitation of Administrative Records for Statistical Purposes, G. Labossière (Statistics Canada) 443

Privacy and Confidentiality Considerations in the Use of Administrative Records for Statistical Purposes?, J.M. Leyes (Statistics Canada) 447

Closing Remarks 451

OPENING REMARKS

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

OPENING REMARKS

I.P. FELLEGI[1]

On behalf of Statistics Canada, I would like to welcome you to the International Symposium on Statistical Uses of Administrative Data, the fourth annual symposium sponsored by Statistics Canada on methodological issues.

This symposium will examine the developments, challenges and issues associated with the acquisition and growing use of administrative records for statistical purposes. I am pleased to note the presence of participants from the academic, private and public sectors of Canada, the United States and several other countries.

I am confident that the information and ideas that will be exchanged over the next three days will benefit participants and observers alike.

Canada's statistical system is heavily dependent upon the use of administrative records for the production of statistics. For a number of administrative files, the principal historic reason for using them, and still the overwhelming reason, is that they provide almost the sole source of essential statistics. Examples are import and export statistics based on custom's declarations, and mortality and demographic statistics based on birth, death and marriage registrations.

Even when administrative records do not constitute the only source of information, they often effectively supplement other data sources. For example, the information on health care obtained from hospital records is greatly enriched, but not replaced, by data on the health and illnesses of the general population obtained through household surveys.

In recent years, there has been a substantial increase in the use of administrative records by Statistics Canada, with the result that in some cases administrative records have replaced surveys. There are several important reasons for this trend. One of the primary reasons was the need to reduce the response burden, particularly on small businesses. The results are impressive, and are largely due to access by Statistics Canada to income tax files. In fact, from 1978 to 1985, the overall response burden of businesses was cut in half, and today 85 percent of small retailers no longer have to complete annual Statistics Canada questionnaires. Data on them are obtained entirely from tax files.

Another important reason for the use of administrative records by Statistics Canada is that they have been generated for other purposes and can be utilized for statistical purposes at marginal cost -- an important consideration at any time, but particularly relevant during periods of budget reductions. Their judicious exploitation -- the emphasis being on both "judicious" and "exploitation" -- is one of the reasons that enabled us to absorb, since 1975, a close to 30 percent budget reduction, while at the same time we were largely able to maintain our product line and, indeed, even expand it in some significant areas.

A third major reason, of our increased reliance on administrative records is their property of providing the basis for the only cost-effective

[1] I.P. Fellegi, Chief Statistician, Statistics Canada, Tunney's Pasture, 26-A, R.H. Coats Bldg. Ottawa, Ontario. K1A 0T6

method of maintaining a comprehensive and unduplicated list of businesses -- which in turn is a cornerstone of our economic statistics. Indeed, it is fair to say that, without the underpinning provided by tax and payroll deduction records, a high quality, well integrated and cost effective economic statistics program could not be maintained.

A fourth impetus for increased reliance on administrative records is the major interest of our clients in subprovincial detail in our statistical output. We established a division to lead the development of this type of information -- and once again administrative records provide the life blood for their work.

Finally, social, health and indeed some economic policy areas require increasing amounts of longitudinal data. This is both difficult and expensive to produce through direct surveys. While we do have some such surveys, we depend disproportionately on administrative records for this kind of information. Exploitation often involves record linkage.

Despite the considerable progress that has been made in utilizing administrative records for statistical purposes since the early 70's, there are a number of issues and challenges which have undoubtedly inhibited their fuller exploitation. These include:

1. Administrative records are frequently held in multiple jurisdictions so the problems of quality, standardization, comparability, consistency, coverage, etc. are intrinsically complex;
2. Administrative data are collected for very specific purposes; hence they may not serve adequately other needs, such as statistics. It is important that a spirit of cooperation be encouraged between administrators and statisticians, based on mutual understanding of each other's problems, objectives and contributions;
3. Not all administrative data are in machine-readable form, a prerequisite for linkages at a micro-level;
4. Concern about negative public perceptions about privacy and the use of administrative records. An important task confronting statisticians is to persuade the public that confidentiality remains fully protected and that we maintain a very rigorous and conservative stance with respect to justifiable privacy concerns: we only engage in record linkage for purposes of producing statistical output -- a class of use which should be clearly separated from enforcement or other administrative "fishing expeditions." Even for statistical purposes we should have rigorous and auditable review procedures to ensure that we only carry out record linkage where the resulting privacy invation is clearly outweighed by the public good from the new statistical information.

We have individually and collectively accumulated considerable information and experience regarding the use of administrative data. The time is ripe to discuss their uses and the associated problems in a public forum. This initiative is clearly supported by the quality of papers... and by the large number of participants who have chosen to participate in this conference. I have no doubt that Statistics Canada will benefit from this symposium through stimulation, the exchange of relevant experience and perhaps through the generation of some important new ideas.

I wish to all of us a productive and pleasant occasion.

SESSION I: INVITED PAPERS

POLICY ISSUES AND ORGANIZATIONAL EXPERIENCE

Chairperson: J. Ryten, Statistics Canada

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

STATISTICAL USES OF ADMINISTRATIVE DATA: ISSUES AND CHALLENGES

G.J. BRACKSTONE[1]

ABSTRACT

Canada's statistical system is heavily dependent on administrative records for the regular production of national and sub-national statistics. After illustrating this dependence, the paper will discuss issues and challenges arising in five different aspects of this topic. Firstly, a review of the different ways in which administrative records can be used will lead into a discussion of whether there should be any policy restrictions on the statistical use of administrative data. Secondly, questions concerning the right of access to administrative records will be raised. Thirdly, the paper will address issues related to the quality or suitability of administrative records for statistical use and their resultant role within the statistical system. Fourthly, the question of how statistical agencies can or should influence the statistical usefulness of administrative records will be raised. Finally, some important issues relating to the public perception of statistical agencies' use of administrative records, and public concerns over privacy and confidentiality will be discussed.

1. INTRODUCTION

Statistical use of administrative data is a topic currently in fashion. What has caused this renewal of interest? — and it is certainly a "renewal" since the origins of official statistics lie in administrative systems. Censuses derive their name from their original purpose of taxation; while much of the early development of official statistics depended on data collected for administrative processes. Statistical censuses have been used widely since the early part of the 19th century, but sample surveys only since the Second World War. Interesting accounts of the development and acceptance of sample surveys as an instrument of official statistics can be found in Hansen, Dalenius and Tepping (1985) and Hansen (1987), while some historical perspective on administrative data use can be found in Bjerve (1985). Why are statisticians now rediscovering administrative records?

The computerization of many administrative programs in the 1960s and 1970s made possible an increased use of the resulting data files for statistical purposes. However, the motivation for the more recent renewed interest in administrative records stems primarily from three factors. Firstly, tightening budgets cause a search for alternatives to the relatively expensive collection costs of statistical surveys and censuses. Secondly, an increasing concern about the burden of statistical enquiries on respondents also leads to a search for alternatives. Thirdly, increasing demands for small area data which

[1] G.J. Brackstone, Assistant Chief Statistician, Informatics and Methodology Field, 26'J' R.H. Coats Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario K1A 0T6.

cannot be obtained from sample surveys motivate the examination of administrative records as a source of such data.

The further technological advances that have facilitated and reduced the cost of manipulating large administrative files in recent years have also encouraged the increased use of administrative records.

In this paper we begin with a discussion of the distinguishing features of administrative records (Section 2) and categorize them by original purpose (Section 3). Many of the papers in this Symposium are concerned with ways in which administrative records are used. Section 4 provides an overview of these methods. A major factor determining how administrative records are used is their quality characteristics, and these are reviewed in Section 5. Section 6 discusses issues of access to administrative records and means of making them more useful for statistical purposes. The important issues of public perceptions of, and concerns about, administrative record use are reviewed in Section 7.

2. DEFINITION OF ADMINISTRATIVE RECORDS

It is worth pausing a moment to consider what exactly we mean by the expression 'administrative data'. The term has become part of statistical parlance — it is the basis of the title of this Symposium — but is rarely defined. There is a general understanding of the distinction between administrative purposes and statistical purposes in the collection of data and this is very important in considerations of privacy and confidentiality. For example, the American Statistical Association's Ad Hoc Committee on Privacy and Confidentiality included the following statements in its report (ASA, 1977):

> "An administrative record is collected and maintained for the purpose of taking action on or controlling actions of an individual person or other entity ... Contrastingly, the statistical record has an entirely different purpose. The statistical purpose is ... to learn the dimensions, trends and relationships of collectives of persons or other entities. The individual identification of a statistical record and its contents is held confidential from all except the persons collecting and compiling the aggregated data. An individual's record is not used to determine any action that affects that individual except through the contribution of the record to statistical aggregates, averages, or measures of relationships. The very essence of statistical analysis is that the identity of individual units ... is immaterial. Individuals should not be identifiable in the output of a statistical system."

Despite this apparently clear distinction between administrative and statistical purposes, when it comes to cataloguing uses of administrative records for statistical purposes within a statistical agency, there remains some ambiguity. We clearly recognize sources such as income tax records and birth registrations as providing administrative records which are subsequently treated like statistical records for purposes of aggregation and analysis. Not all cases are so clear.

There are several different ways in which administrative records might become available to the statistician. For example:

(a) They may be collected by the administrative agency and provided to the statistical agency as a file of individual records;

(b) They may be collected by the administrative agency and provided to the statistical agency in an aggregated form;

(c) The statistical agency may conduct a survey (or census) of local administrative agencies (e.g. municipalities, school boards) seeking data about the individual units (e.g. dwellings, students) these agencies administer. These data may be provided in the form of individual records or aggregated for each local jurisdiction.

While categories (a) and (b) may be instantly recognizable as examples of administrative record use, category (c) is ambiguous. An administrative process collects data from the individual units; a statistical process (survey or census) brings the data into the statistical agency. If the individual unit records reach the statistical agency we may regard this as administrative record use. However, if they arrive in aggregated form, how does this differ from, say, a survey of businesses enquiring about their workforce — data presumably assembled within the business by an administrative process? Is the distinction between micro-data and aggregated data the key element? Is the question of whether the administrative agency is in the public sector or not important?

It may be less important to have a watertight definition than to have an understanding of the features that distinguish administrative data from data from statistical sources in the context of statistical use. We suggest the following features:

(i) the agent that supplies the data to the statistical agency and the unit to which the data relate are different (in contrast to most statistical surveys);

(ii) the data were originally collected for a definite non-statistical purpose that might affect the treatment of the source unit;

(iii) complete (100%) coverage of the target population is the aim;

(iv) control of the methods by which the administrative data are collected and processed rests with the administrative agency.

Each of these features affects the characteristics of administrative records and has implications for the way administrative records are used within a statistical system.

3. TYPES OF ADMINISTRATIVE RECORDS

Even without an explicit definition of administrative records, one can nevertheless categorize them according to their initial purpose. The purpose for which administrative records were originally assembled can have a profound effect on their usefulness for statistical purposes in terms of their coverage, content and the accuracy of the specific data items they include. Among the main categories that can be distinguished are the following:

(i) Records maintained to regulate the flow of goods and people across borders (e.g. imports, exports, immigration)

(ii) Records arising from legal requirements to register particular events (e.g. vital statistics, business incorporations, licensing)

(iii) Records needed to administer benefits or obligations (e.g. taxation, unemployment insurance, health insurance, family allowance)

(iv) Records needed to administer public institutions (e.g. schools, hospitals, prisons)

(v) Records arising from the government regulation of industry (e.g. transportation, banking, broadcasting)

(vi) Records arising from the provision of utilities (e.g. electricity, phone, water services).

A second distinction is between those administered nationally (usually by the Federal Government) and those administered sub-nationally (e.g. by provinces or municipalities). For the latter to be useful nationally, agreement between jurisdictions is required on items such as definitions, standards, record formats, and procedures. Such agreement is not always easy to achieve, particularly in domains that are constitutionally within provincial jurisdictions.

There are other dimensions in which administrative records can be categorized: by collection process (personal interview, mail, observation, etc.); by storage medium (paper or computer file); by accessibility (see Section 6).

4. USES OF ADMINISTRATIVE RECORDS

Much of this Symposium is concerned with ways in which administrative records may be used for statistical purposes. These uses can be divided into five broad categories. Most statistical applications of administrative records fall into one of these categories or represent combinations or variations of these uses.

(i) Direct tabulation

This includes the counting of units in files, cross-classification by attribute, and the aggregation of quantitative variables associated with each unit. Statistics on vital events and on external trade are important examples. Other examples include the publication of monthly counts of unemployment insurance claimants or beneficiaries by province, age, sex and length and type of benefit, or annual summaries of income distributions for each county based on the personal income tax file.

(ii) Substitution for survey data

This category covers cases where some strata are not surveyed but administrative data are used instead. The use of tax data for small businesses is such a case in Statistics Canada. One might extend this category to include cases where administrative data are used in an imputation process to handle survey non-response.

(iii) Indirect estimation

This category includes cases where data from administrative records comprise one of the inputs into an estimation process. For example, individual tax returns for the same taxfiler are linked from one year to the next in order to produce partial estimates of migration which can be weighted up with reference to census-based benchmarks. These estimates of migration then feed into Statistics Canada's population estimation program (which also makes use of administrative data on births, deaths and immigration). A second example would be the use of administrative data as a source of auxiliary variables in ratio or regression estimation.

Also within this category are uses that involve the linkage of different administrative or statistical files to produce estimates. For example, the linkage of the death register with files of individuals exposed to particular hazards in order to estimate differential mortality rates, or the linkage of records from tax files, unemployment insurance files, and manpower training files in order to analyse labour market attachment and adjustment.

(iv) Survey frames

In this category we include the use of administrative records to create, supplement or update frames to be used for censuses or surveys. A primary example is the use of payroll deduction information submitted by employers to Revenue Canada. The questionnaire which has to be completed by new payroll deduction account holders is a valuable means of identifying new businesses or changes in the structure of existing ones. Although in Canada we do not at present have a register of housing units, a second example would be the use of building permits or new telephone or electricity connections as signals of possible new housing units.

(v) Survey evaluation

This category covers the use of administrative records for checking, validating or evaluating survey-derived data. This may be done either at the individual unit level, or at an aggregate level. Several Census evaluation studies in the past have used immigration and taxation records to evaluate Census questions on immigration and income respectively, while family allowance records have been used in checking the census coverage of children.

Many applications involve the combination of administrative record use with other methods of data collection such as sample surveys and censuses. For example, the essence of Statistics Canada's business survey redesign project is the combined use of administrative records and surveys for both frame maintenance and economic data collection (Colledge, 1987). Clearly the program of population estimates involves a combination of censuses and administrative records, with sample surveys also being used for the evaluation of census coverage. Examples of integrated approaches in Statistics Canada and elsewhere will be covered in other sessions of this Symposium.

The ways in which administrative records can be used are largely a function of their content and quality characteristics. We will turn next to a consideration of quality issues related to administrative records.

5. QUALITY OF ADMINISTRATIVE RECORDS

The suitability of administrative records for a particular use will depend on a variety of factors of which the following may be most important:

(i) the intended coverage of the administrative system;

(ii) the content of, or the variables included in, the administrative system, and the concepts and definitions underlying them;

(iii) the quality with which data are reported and processed in the administrative system;

(iv) the timeliness with which the data are available for statistical use.

The first two of these factors are related to the purpose of the administrative program. The target population, the information items collected, and the definitions used, are generally dictated by the administrative purpose. There is an important issue concerned with the extent to which statistical needs might influence the coverage or content of administrative records. We will return to it in Section 6.

The third factor reflects the degree to which the administrative system succeeds in obtaining the coverage and content it seeks, and is of prime importance in assessing the statistical usefulness of an administrative source. One cannot generalize about the

quality of administrative records; it is necessary to examine the quality characteristics of individual administrative record sets.

The fourth factor, timeliness, is an important consideration when building the use of administrative records into statistical processes. The elapsed time between the reference period and the availability of a file for statistical purposes, and the confidence one can have in regularly receiving the file on time, both have to be taken into account.

The following points summarize some features of administrative record systems that may influence their usefulness for statistical purposes:

1. What incentive is there for individual units to be registered in the administrative system? Programs that lead to benefits for registrants (e.g. family allowance, health insurance) should result in very complete coverage of eligible persons. There may even be problems of overcoverage if the mechanism for removing those who cease to qualify is deficient. Programs not seen as beneficial to registrants (e.g. taxation, licensing for some purposes) may result in less than complete coverage.

2. Administrative records do not usually represent a rich source of cross-classified data because they generally collect only the limited set of variables required for the administration of a program. For example, the income tax file contains detailed income data but only limited other data — age and sex but not education level or industry.

3. Since the concepts and definitions used in administrative systems are defined to meet program needs, they will not necessarily coincide with those required for social or economic analysis. Use of an income tax file as a source of data on the working age population has to be tempered by the fact that not all persons in this age group are taxfilers. Furthermore, some of the groups with low coverage in the tax file are precisely those of key analytical importance (e.g. low income earners, the elderly). A second familiar example concerns persons receiving unemployment insurance benefits who are not synonymous with persons defined as unemployed by international statistical standards.

4. The coverage and content of administrative records can be subject to discontinuities resulting from changes to laws, regulations or administrative practice. Estimation procedures that involve calibrating administrative data to other data sources will be disrupted by such changes. For example, the introduction of Child Tax Credits in the late 1970's led to a sudden increase in the coverage of the individual tax file.

5. Administrative records are a potentially valuable source of small area data because of their census-like nature. However, such use requires a precise geographic location code to be present on each record. Postal codes may serve this purpose where address information is present. Care has to be taken that the address given reflects the appropriate location. For example, tax files may show a discounter's address rather than the address of the taxfiler.

6. Quality assurance procedures applied to data in the administrative system may be very tight for variables critical to program administration but much less stringent for other variables. As mentioned above, geographic identifiers may be imprecise.

7. Administrative files about individuals will often not identify families or households. Combining individuals into families or households may be possible in some cases, but often the required matching information will not be available.

These points may have tended to emphasize the deficiencies in administrative records rather than their strengths. They have to be considered in conjunction with the strengths and weaknesses of censuses and sample surveys in the production of statistical information. In many situations the optimum statistical program uses a combination of these different data sources in an integrated fashion. The United Nations Statistical Handbook summarized well [4]:

> "A balanced programme for the improvement of national statistics involves the use of censuses, sample surveys and administrative records. In the long run these three sources of statistics are complementary to a considerable extent; each has some advantages and suffers from some limitations. Therefore the full development of one source does not render the other two sources superfluous."

6. ACCESSING AND INFLUENCING ADMINISTRATIVE SYSTEMS

A further set of important issues relate to the statistician's access to, and influence on, administrative record systems. It will be clear by the end of this Symposium, if not already, how heavily dependent many statistical agencies, including Statistics Canada, are on administrative records. Therefore, measures to ensure continuing or expanded access to such records, as well as measures to make them more useful for statistical purposes are needed.

Access

Legal authority for access to administrative records is required. For Statistics Canada it is provided by Section 12 of the Statistics Act (1971):

> A person having the custody or charge of any documents or records that are maintained in any department or in any municipal office, corporation, business or organization, from which information sought in respect of the objects of or correction thereof, shall grant access thereto for those purposes to a person authorized by the Chief Statistician to obtain such information or such aid in the completion or correction of such information.

While this provision appears to give fairly broad access rights, it is not without limitations. In some cases, legislation governing the administrative process places restrictions on access or secondary use of the administrative data. This leads to a confrontation of legislation that will at best delay the negotiation of access. In some cases, access for statistical purposes is specifically permitted.

Enabling legislation is a necessary but not sufficient condition for the productive utilization of administrative records. A co-operative approach to the development and utilization of administrative records for statistical purposes is likely to be far more effective in obtaining access to administrative records than an approach involving legal arguments and sanctions. Indeed, once access is obtained, the subsequent step of influencing design or procedures is only achievable if there is a spirit of co-operation between the administrative and statistical agencies.

In Canada, access to administrative records by Statistics Canada is strictly a one-way street. Individual micro-data are provided from the administrative agency to the statistical agency, but only confidentiality-protected aggregate data can flow back. The only exception to this rule is the case where the administrative agency depends on the statistical agency to organize, format, edit, process, or restructure its records, and a version of the original microdata is passed back to the supplying agency.

Influencing change

We have already alluded to the potential impact of changes in administrative regulations or practices on resulting statistics. Discontinuities in time series based on administrative records can be caused by simple changes in the coverage of a program, the introduction of an incentive to join or leave a program, or procedural changes that affect quality or completeness of records. Thus the statistical agency has to guard against, and react to, externally imposed changes.

There are other kinds of changes that the statistical agency might like to see implemented. A frequent frustration of the statistician trying to use administrative records is the feeling that the administrative records could be so much more useful if only relatively minor changes were made. For example, the addition of an extra question, the use of a different concept, the coverage of an additional sub-group, or the introduction of a quality check might significantly enhance the statistical value of the records. On the other hand, why should the administrative agency contemplate changes not required for the primary administrative purpose, changes which would probably in some measure add to the cost and complexity of the administrative process?

The challenge for a statistical agency is to persuade the administrators that the benefits from such a change outweigh any additional administrative costs. This is made harder to the extent that the benefits do not accrue to the department responsible for the administrative system, but to separate policy-making departments and other statistical users.

This is one area where a decentralized statistical system might have advantages over a centralized one, such as Canada's. One might expect it to be easier for statisticians working within a department or ministry to influence departmental systems than those located in a central statistical agency. Perhaps, some of our visitors from the U.S. or elsewhere will be able to shed some light on that question.

In the context of a centralized statistical system, there are some mechanisms that might serve to give added weight to statistical considerations when administrative systems are being designed or redesigned. Some of these mechanisms are bilateral, others are government-wide.

1. Bilateral committees at a senior level to review and discuss issues of mutual interest, including problems related to the supply of administrative data;
2. Feedback of statistical data to the administrative agency to demonstrate both usefulness of the data and, perhaps, weaknesses arising from administrative practices;
3. Provision of technical advice or services in support of the administrative agency's own statistical activities;
4. A Government information collection policy that requires, for example, any data collection activity plan (statistical or administrative) to be reviewed by a central agency;
5. Statistical planning in the form of a requirement that each new program proposal include a plan for acquiring the statistical information needed to monitor and evaluate the program;
6. Promotion of the use of standard statistical definitions (e.g. family, business establishment, unemployed) in administrative systems;
7. Audits that identify the use of administrative records as a cost efficient alternative to other means of acquiring information;

8. Political instruction to make greater use of particular administrative systems or seek alternatives to survey-taking;

9. Removal of legislative impediments to access or use of administrative records for statistical purposes.

Statistics Canada's experience in dealing with other federal government departments has been most successful in cases where close bilateral arrangements have been developed. The introduction of senior bilateral committees in the early 1980s was supportive of such arrangements, and in some cases instrumental in creating them. Government-wide measures such as information management and statistical planning have been less successful in facilitating administrative record use. Government audits and cabinet directives have provided impetus to activities aimed at increasing administrative data use, but the increased use itself is again dependent upon close working relationships with particular departments. While it is convenient to characterize the statistical agency as the progressive agency trying to break down unreasonable barriers to administrative data use, it must also be recognized that there may be inertia to the associated changes within the statistical agency itself. Staff whose careers have been based on survey design and survey-taking may need convincing that budgetary restrictions and data needs now necessitate combining these with other approaches.

Since the above comments have focussed on federally administered systems, we will add a few words about provincial records. While some of the above measures apply equally to provincially administered records, the fundamental problem in dealing with subnational jurisdictions is that of adherence to common definitions and standards. Differing provincial needs and priorities, facilitated by increasing technological capacity, will lead to divergent administrative systems in the absence of any centralizing force. Statistics Canada has used a variety of mechanisms in the past in attempts to encourage conformity, but with only mixed success. As with federal government custodians of administrative records, mutual benefit has to be the major incentive to conformity. Federal-provincial committees exist in several subject areas. The Vital Statistics Council consisting of provincial registrars of vital events and representatives of Statistics Canada is a successful and long-standing example. Such committees have developed and monitored conventions for reporting certain data items in the past. For example, the framework for municipal finance reporting was developed as a result of federal-provincial meetings on municipal financial statistics.

7. PUBLIC PERCEPTIONS

Our final, and perhaps most important, set of issues concerns public perceptions of administrative record use. We refer to the negative images conjured up by notions of vast databanks constructed by linking records and through which the characteristics and actions of individuals can be traced. Debates over whether data collected for one purpose should be used for another purpose fall in this category. At the heart of these concerns is the protection of privacy and the need to ensure that there is management and control of the use of personal information.

On the other hand, there are positive public perceptions to be stressed in the use of administrative records for statistical purposes. Reduction in respondent burden and reduction in government expenses are positive effects of administrative record use.

There are three main conditions that are probably necessary if the use of administrative data for statistical purposes is to gain acceptance.

Firstly, there has to be a recognition that statistical work is a legitimate secondary use of administrative records. The Privacy Act contains this recognition under certain

conditions. However, it is doubtful whether the average citizen appreciates the distinction between statistical use, where the identity of the individual record is of no lasting interest, and administrative use where the essence of the individual record is the particular unit to which it relates. It would be easier to explain and utilize this distinction if we could state unequivocally that identifiers are never needed for statistical purposes. Unfortunately this is not the case. Several legitimate statistical techniques do require identifiers in intermediate data manipulations. These techniques all involve some form of matching data from different files or different occasions, and identification is required to ensure that the correct records are matched. Once the matching has been accomplished the records can be anonymized provided no subsequent linkage is planned. Examples include the requirement for names in a population census to ensure coverage and permit coverage measurement, longitudinal studies using administrative records, epidemiological investigations, and evaluation studies to check survey responses against administrative sources. Explaining why identifiers are needed when identity is of no interest is an interesting challenge facing the statistical agency.

Secondly, there should be demonstrable control over record linkage activities. Statistics Canada has developed a policy on record linkage. The Policy recognizes the benefits of record linkage to statistical programs as well as the concerns that unconstrained record linkage create, and indicates that decisions on record linkage activities require judgement to balance these two factors.

Essentially the Policy states that Statistics Canada will engage in record linkage activities if all of the following conditions are satisfied:

- the record linkage is for statistical or research purposes consistent with the mandate of Statistics Canada;
- the products of the linkage will be released only in accordance with the confidentiality provisions of the Statistics Act;
- the linkage has demonstrable cost or respondent burden savings over other alternatives, or is the only feasible option;
- the linkage will not be used for purposes that can be detrimental to the individuals involved, and the benefits of the linkage are clearly in the public interest;
- the linkage activity is judged not to jeopardize the future conduct of Statistics Canada's programs;
- the linkage satisfies a prescribed review and approval process.

The review and approval process requires documentaiton of each record linkage proposal, review and recommendation by Statistics Canada's internal Confidentiality and Legislation Committee, and approval by the Chief Statistician and the Minister.

This process allows record linkage to take place where it is well justified and in the public interest but ensures that record linkage is not undertaken lightly.

Thirdly, confidentiality of data from administrative records is essential. This includes both the protection of individual records and the checking of aggregates to ensure there is no inadvertent disclosure in published tables. Confidentiality is, of course, a requirement for all microdata in a statistical agency, not just that from administrative records. The one-way flow of microdata from the administrative agency to the statistical agency should be stressed.

In addition to these three main points, there are other measures that can be taken to counteract a potentially negative public perception of administrative record use. Public communication can stress confidentiality and the benefits in terms of burden reduction

and costs. Statistical purposes should be included in the statements of purpose that are required by the Privacy Act in any collection of personal information.

While the above points represent some specific measures that can be taken to avoid or respond to public reaction to the use of administrative records, ultimately the statistical agency must have strong political support for this kind of activity. The political credit to be gained from demonstrated reductions in costs and respondent burden, coupled with strong political assurances of the protection of individual data, provide a firm platform for politicians to dispel public concern over the use of administrative records for statistical purposes. At the same time they must immediately and unambiguously confront and correct any suggestion that statistical records be used for administrative purposes.

8. CONCLUSION

Administrative records are and will continue to be an increasingly important source of statistical data. The relative strengths and weaknesses of data derived from administrative systems, in terms of cost, coverage, quality, relevance and timeliness, in comparison to census or survey based data, dictate the manner in which these sources of data are most effectively used. Current uses of administrative records include direct tabulation, indirect estimation, substitution for survey responses, frame construction and maintenance, and data evaluation. These uses now permeate most statistical programs and can be expected to extend even further in the future.

In Canada, administrative records have become part of the fabric of our statistical system. Their use has been one of the means by which Statistics Canada has been able to maintain its programs in the face of declining budgets. In the process, respondent burden has been reduced and new, or more frequent, data series have become available. Since we do not have administrative registers as such, considerable attention has been paid to issues of coverage and the joint use of both administrative and survey based data to ensure valid estimation of universe totals. The use of record linkage techniques, though requiring careful controls, has proven to be very valuable, particularly for business data, longitudinal labour market studies, and epidemiological work.

With the growing use of administrative records, statistical agencies are becoming increasingly dependent upon other agencies for the uninterrupted flow of input data to their statistical programs. Whatever the legislative and policy environment in which the statistical agency operates, the establishment of close co-operative arrangements with supplying agencies is crucial. The ability of the statistical agency to influence the design or redesign of administrative systems rests on a mutual understanding of the requirements of the two agencies. Establishment of a government-wide policy or principle that the statistical agency should have a voice in decisions regarding the design of administrative systems, or more generally, in proposals for meeting the statistical needs of new programs, can help the statistical agency in this regard, but is no substitute for the fostering of close co-operation with administrative agencies.

The variety of issues arising in the use of administrative records are well reflected in the sessions of this Symposium. In addition to overall policy issues addressed in this session, there will be sessions on record linkage, quality of data, and the use of administrative data in conjunction with or as a partial substitute for census or survey data. The Panel discussion on Wednesday will focus on privacy and public perception issues.

There is such a richness of data in administrative systems that the real challenge for statisticians is to influence and harness these data sources so that they become a fully integrated part of the statistical system. This has happened in some domains in some countries, but there is much still to do and many benefits still to gain.

REFERENCES

American Statistical Association (1977). Report of the Ad Hoc Committee on Privacy and Confidentiality. *The American Statistician.* 31, 59-78.

Bjerve, P. J. (1985). "International Trends in Official Statistics" in A Celebration of Statistics, *The ISI Centenary Volume.* Ed. Atkinson, A.C. and Fienberg, S.E.

Colledge, M.J. (1987). "The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada" Proceedings of the Third Annual Research Conference, March 1987, U.S. Bureau of the Census, Washington, D.C.

Hansen, M. "Some History and Reminiscences on Survey Sampling". *Statistical Science,* May 1987, Vol. 2 No. 2, 180-190.

Hansen, M. H., Dalenius, T., and Tepping, B. J. (1985). "The Development of Sample Surveys of Finite Populations" in A Celebration of Statistics, *The ISI Centenary Volume.* Ed. Atkinson, A.C. and Fienberg, S.E.

Statistics Act (1971). Statutes of Canada 1970-71-72, c.15.

United Nations (1980). "Handbook of Statistical Organization", Volume 1, A Study on the organization of national statistical services and related management issues.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

EUROPEAN EXPERIENCE OF USING ADMINISTRATIVE DATA FOR CENSUSES OF POPULATION: THE POLICY ISSUES THAT MUST BE ADDRESSED

PHILIP REDFERN[1]

ABSTRACT

The experience of the four Nordic countries illustrates the advantages and disadvantages of a register-based census of population and points to ways in which the disadvantages can be contained. Other countries see major obstacles to a register-based census: the lack of data systems of the kind and quality needed; and public concern about privacy and the power of the State. These issues go far beyond statistics; they concern policy and administration. The paper looks at the situation in two countries, the United Kingdom and Australia. In the United Kingdom past initiatives aimed at population registration in peacetime foundered and the present environment is hostile to any new initiative. But the government is going ahead with a controversial reform of local taxation that involves setting up new registers. In Australia the government tabled a Bill to introduce identity cards and an associated register, and advanced clearcut political arguments to support it; the Bill was later withdrawn. The paper concludes that the issues involved in reforming data systems deserve to be fully discussed and gives reasons why statisticians should take a leading part in the debate.

1. INTRODUCTION

This paper has its origin in a study of alternative approaches to the census of population that I carried out for the Statistical Office of the European Communities (Redfern 1987). The study examined the experiences of the 12 member countries of the EEC together with Canada, Sweden and the United States. The study found that sample surveys can complement, but cannot replace, a 100 per cent census, because they do not provide reliable statistics for small areas. An important example of samples complementing a 100 per cent enumeration is the short form/long form censuses of Canada and the US. A sample survey complementing 100 per cent data from registers is in prospect in Norway (Section 3.3).

Registers that contain addresses give figures for small areas; and, if the registers cover the census topics reliably (in terms of definitions, coverage, accuracy and timeliness) and can be linked, it is possible to create a record for each individual akin to his census return and so to conduct a register-based census: in essence administrative

[1] Philip Redfern, United Kingdom, 17 Fulwith Close, Harrogate, North Yorkshire, England, HG2 8HP.

data are being recycled for statistical purposes. The pressure of costs and the burden of formfilling in the traditional census have persuaded the Nordic countries (Denmark, Finland, Norway and Sweden) to adopt this approach in whole or in part.

Though administrative data can support a **conventional** census in a variety of ways (Redfern 1987), it is their use in a **register-based** census that provides the first main theme of this paper. Section 2 describes the registers that are needed as a base for a census and Section 3 identifies the similarities and differences between the four Nordic countries in their approaches to this kind of census. Section 4 then considers the obstacles that other countries would face if they were to upgrade their record systems so as to make a register-based census feasible, and recognises that the issues raised concern administration and policy more than statistics.

It is these wider issues that provide the second main theme of the paper. Section 5 looks in more detail at a country in which, for reasons of policy and ideology, administrative records are not coordinated through a population register: the United Kingdom. Section 6 described a recent initiative in Australia to improve administrative records. Finally Section 7 summarises the political arguments for and against coordinating administrative records through population registers and puts the case for statisticians taking a leading part in debate on the subject.

2. THE REGISTERS NEEDED AS A BASE FOR THE CENSUS

2.1 Population Registers

The essential starting point for a register-based census is a population register that includes personal reference numbers and addresses. The personal numbers must be in one to one correspondence with the members of the population. To keep the register up-to-date the citizen is obliged to notify changes. The personal numbers are also recorded in the files of the various administrative agencies, and so can be used to link records for statistical purposes.

Population registration serves essentially administrative ends. It is an efficient way of organising the many dealings between public authorities, both central and local, and the individual citizen: for example taxes, social security, publicly-provided health services and electoral registration. To work effectively, population registration should serve a wide range of administrative activities, so that opportunities for updating and correction are frequent and the citizen becomes used to quoting his personal number.

The key to the system is the central population register which records identifying information about each person (name, place and date of birth, date of immigration, marital status, and possibly items like parentage and citizenship) and his permanent reference number. In most countries the central population register includes up-to-date addresses, though the French **Répertoire National d'Identification des Personnes Physiques** does not. The basic administrative function of the central register is to act as reference point for administrative agencies which can check the identities of the individuals that they are dealing with and, as necessary, can correct or record the personal reference numbers in their own files.

2.2 Other Key Registers in a Register-Based Census

A register-based census of population and housing makes use of registers of other kinds of units than persons. The most important are a central register of housing and a central register of business enterprises and establishments (workplaces). Provided the housing register identifies each housing unit (and not just the building or the address) with a code that also appears as part of the address in the population register, then data on the

housing unit in the housing register can be associated with data on the occupants in the population register: the two registers can be linked. Similarly a register recording each person's employer and workplace can be linked to a central register of enterprises and establishments to show the person's industry, commuting journey, etc.

3. CENSUSES IN THE NORDIC COUNTRIES

The four Nordic countries have well-developed population registers of the kind described in section 2.1. They have constructed, or propose to construct, central registers of building and housing to serve mainly administrative purposes. This section of the paper outlines the census of each country in turn and then summarises the directions in which Nordic census-taking is developing.

3.1 Denmark

Denmark is the only Nordic country — and I believe the only European country — to have switched completely from the conventional census to a register-based census. The switch was made in little more than a decade. The central population register with personal reference numbers was created in 1968 for administrative purposes, and a register-based census of population (but not housing) followed in 1976. A central register of buildings and dwellings was created in 1977, again mainly for administrative purposes, and a register-based census of population and housing followed in 1981. Another significant step in 1979-80 was to extend the return in which employers report each employee's earnings to the tax authorities: employers with more than one workplace added each employee's workplace to the return. This was done purely for statistical purposes and the statistical office has had to make a considerable effort to secure a good response.

The registers held by Danmarks Statistik for statistical purposes, numbering some 37, provide annual or more frequent statistics of population, employment, commuting, income, housing and construction for municipalities and sometimes smaller areas. But, because of the cost, analysis on the scale of a census takes place much less frequently: the next after the 1981 census will take place in 1991 and even that may be on a lesser scale than 1981.

The transition to a register-based census has been facilitated by the reorganisation of the Danish central statistical office in 1966. Danmarks Statistik was given a measure of independence of the central government, which could help to reassure the public on confidentiality. It was given powers to demand, and to use for statistical purposes, data held by public authorities for administrative purposes, and to participate in the construction of registers containing such data.

The problems that Danmarks Statistik now faces concern mainly the quality and timeliness of data, both of which depend on the efficiency of administrative procedures. Thus the slowness in compiling tax authorities' files — which provide data on industry, occupation, journey to work and income — delayed analysis of these topics in the 1981 census until summer 1983; and it is expected that statistics on the labour force will continue to lag at least a year behind the reference year to which they relate. Reliable data on occupation are particularly difficult to obtain because the topic is of little administrative interest; a main source is the information given by the taxpayer on his annual tax return. Despite problems of these kinds Danmarks Statistik takes the view that the register-based census has come to stay in Denmark because of the savings in cost and in burden on the public (Jensen 1983).

3.2 Finland

Register-based censuses have a long history in Finland. In the 1600s the parish registers recorded everyone over the age of 12 living in the parish, and in 1749 figures of the total population were compiled analysed by age, sex, marital status and social class: one of the first-ever register-based censuses? Later censuses followed this pattern. The censuses of 1950 and 1960 adopted the conventional method of collecting the information through questionnaires. But beginning with the 1970 census an increasing range of data has been extracted from registers. In the mid-decade census of 1985 the questionnaire asked only about economic topics: type of activity (if any) and occupational status, employer and workplace, occupation, and number of months worked in the past year. Data on housing were taken from the register of buildings and dwellings that had been created from 1980 census data and is updated with information from the municipalities.

The 1985 census was planned to cost a little under the equivalent of 1 US dollar per person, or only a **quarter** of the cost of the 1980 census in real terms though covering the same range of variables. Factors that helped to make this possible included: mail-out of questionnaires preprinted with data on workplace (from the 1980 census) and occupation (from the central population register) — to be corrected by the respondent if necessary; mail-back to the central office with no local field organisation; only one reminder, with no follow-up of the 3.7 per cent of forms which were not mailed back or were mailed back incomplete; and imputation of missing data, where possible, using a variety of registers, one of which was pension records in respect of private sector employment. The final response rate to the questionnaire was 97.4 per cent, and by imputing missing data a final coverage of 98.6 per cent was achieved. Another reason for the low cost of the census is that part of the cost and burden has been transferred to the registration systems, including the annual field checks on the population registers by means of forms issued to each household/dwelling and quinquennial checks on the register of buildings and dwellings by means of forms sent to owners and occupiers.

Comparisons between the 1980 census responses and register data on economic variables have been regarded as encouraging. This, and the methods developed in the 1985 census to impute the economic characteristics of non-respondents, open up the possibility that the 1990 Finnish census might be wholly register-based. To fill one gap in register data, employers with more than one workplace will in future make a return of each employee's workplace (Laihonen and Myrskylä 1987; Heinonen and Laihonen 1987).

3.3 Norway

The 1980 census of Norway was to a substantial extent register-based. It took data on basic demographic topics, income and completed education (other than education abroad) from registers. These data were complemented by means of a mail-out mail-back questionnaire to each person aged 16 and over on economic topics, education abroad, country of birth, religious affiliation and housing. All persons in the same household were to return their forms, together with one housing form, in the same envelope, thus defining the composition of the household for census purposes.

For several reasons it is not feasible to switch to an entirely register-based census in 1990. First, register data on some important census variables do not conform to desirable statistical definitions or are not of sufficient quality for census purposes (this applies for example to industry); and register data for other variables do not exist (for example occupation). Second, the development of the register of land property, addresses and buildings (the "GAB" register), begun in 1983, is unlikely to be far enough advanced by 1990 to provide housing data for the census. Third, because the link between the GAB register and the population registers is the address, it is not possible to identify household composition or to associate housing characteristics with personal characteristics when two or more housing units have the same address.

In the 1990 census data from registers will again be used for basic demographic topics, income and completed education (other than education abroad). A method is being developed for converting register data on most of the economic variables to statistically-desirable definitions by reference to the results of an enquiry addressed to a 10 per cent sample of persons aged 16 and above (100 per cent in municipalities with populations under 6,000). The register data for a sub-population would be adjusted in part using sample data for the sub-population and in part using sample data for a wider population — a procedure that would partially eliminate the bias in the register data. The sample enquiry would be the only census source for topics for which no register data exist, including occupation and probably housing and household composition.

This approach — the use of registers plus a 10 per cent sample enquiry — is estimated to cost 60 per cent of the cost of a census on 1980 lines. The penalties would be the sampling variance, which would be greatest for topics for which no register data exist, and also some bias in the case of topics for which register data exist but are not of the quality needed for census purposes (Johansen 1987).

3.4 Sweden

Over the past two decades the balance of the Swedish census has changed: in 1970 most of the data came from questionnaires and a few from registers, but in 1985 the position was reversed. In 1985 the mail-out mail-back questionnaire to each person aged 16 and over (or married couple) asked only (1) whether the person was economically active in a specified week and, if so, the occupation, (2) the household composition — a list of the adults who live in the dwelling and (3) housing questions. It was possible to omit questions asked in the preceding census on the name of the enterprise at which the person was employed, the workplace and the industry, because from 1985 the annual returns that employers make to the tax authorities giving each employee's earnings were extended to show the employee's workplace. But the topic hours of work was dropped from the 1985 census when employers resisted the proposal to include this too on the annual returns.

After the 1980 census a study had been made of the steps that would have to be taken if the 1985 census were to be wholly register-based. The steps included:

1) The use of data on occupation from the forms on which employed persons report changes in income to the national insurance offices.

2) The creation of a register of household composition, which would be updated by asking for more information when a person moved house.

3) The creation of a register of buildings that contain housing units, to be updated by the municipalities.

4) The creation of a register of completed education, to be updated with information from educational institutions on new graduations.

But, as already noted, a questionnaire was retained in the 1985 census mainly because of doubts about the quality of information that could be obtained from registers on occupation, household composition and housing. Of the proposed new registers only the register of completed education is as yet under construction. But a committee is studying the possibility that the record of a person's address in the population registers should include the housing unit and not just the property — an essential step in linking population registers to housing registers.

A Parliamentary Commission is reviewing the 1985 census, particularly aspects concerning privacy and confidentiality. Its findings will be one of the factors shaping the 1990 census. A final report is expected in 1988.

3.5 Summary of Nordic Census-Taking

The four Nordic countries are developing their censuses along different paths but there are many features in common:

1) All have as a starting point accurate registers of population which give regular and reliable statistics of population for small areas.

2) All wish to maximise the use of information in other registers and to minimise the burden of formfilling on the public. All are striving to contain or reduce costs.

3) All recognise the problems of definition, quality and timeliness of the information in registers, particularly for economic topics. Employers' returns are being extended to give information on each person's workplace, and hence on industry — though extensions for purely statistical purposes are unwelcome and may yield data that are of poor quality. Register data on occupation are generally unreliable. And data on some topics, such as method of travel to work, do not exist in any register.

4) Registers of buildings and houses have been created or are proposed. But it is difficult to keep the registers up-to-date, whether by using information available to the municipalities or by collecting information directly from owners. In some countries the registers need to be further refined to identify each housing unit in a way that permits a link with the address information in the population registers. Another problem is how to get data on household composition from registers if, as in Sweden, the household is **not** defined as all the occupants of the housing unit.

All four countries appear ready to sacrifice something in the quality of the census results in order to cut costs and the burden on the public. But they differ in their approaches. Denmark has gone the farthest by abandoning the census questionnaire. Because of doubts on the quality of some register data, particularly on economic topics, the 1985 censuses in Finland and Sweden retained a limited questionnaire, and the responses were linked to demographic and other data taken from registers. But the possibility is foreseen of making the 1990 census of Finland wholly register-based. In Norway, where there was no mid-decade census, the 1990 census is expected to retain a questionnaire on at least economic topics but, to reduce costs, the questionnaire may be sent only to a 10 per cent sample of persons; where register data for economic topics exist, though imperfect, they could be converted to statistically-desirable definitions by reference to the sample data. A valuable account of Swedish experience of using registers as a census source has been given by Johansson (1987).

4. THE FEASIBILITY OF A REGISTER-BASED CENSUS IN OTHER COUNTRIES

The two main forces that have driven the Nordic countries towards a register-based census — the need to cut costs and the burden of formfilling — have been strongly at work elsewhere. They show for example in a halt, and sometimes a reversal, of the pre-1980 trend to longer census questionnaires.

A new and disturbing feature, public protest, disrupted the census in two countries. In the Netherlands the plans for a 1981 census were abandoned. The census in the Federal Republic of Germany planned for 1983 had to be postponed to 1987 because of more stringent conditions on confidentiality laid down by the Constitutional Court, and even then there was some non-cooperation. No country can feel itself secure against this kind of challenge. But a register-based census is less likely to be sabotaged provided it does not have to be supplemented by a questionnaire. This is because there is no occasion (Census Day) when everyone is faced with a questionnaire and the protests of a minority can be fanned into large-scale opposition.

If the register-based census is so much cheaper with less burden on the public and less risk of sabotage, why do so few countries see it as a viable methodology? There are three main reasons. First, for some topics, particularly economic topics, administrative data may be of poorer quality than data collected through questionnaires; and for other topics no administrative data exist. The Nordic countries recognise these shortcomings, and so some have retained a questionnaire and linked the responses to the data from registers (Section 3.5).

Second, many countries do not possess the necessary data systems of the kind described in Section 2. For example, local population registers may exist but without a central population register, as in the Federal Republic of Germany, Greece and Italy. The population registers may not be up-to-date and indeed some countries rely heavily on the canvass for a conventional census of population to update the registers (Italy and Spain). Outside the Nordic group, the Benelux countries have, or are likely soon to have, the data infrastructure needed for a register-based census.

The third main obstacle to a register-based census follows from the second. If the data systems have to be radically improved — and particularly if there has to be wider use of personal numbers and a new obligation to notify each change of address — opposition may be expected from politicians and the public on grounds of privacy and erosion of freedom. There may be doubts too whether the public would cooperate in the bureaucratic disciplines of a good register system. In addition, even when the necessary data infrastructure is in place, its use for record linkage for census or other statistical purposes could be sensitive. These are important issues but they go far beyond statistics. They concern policy and administration. They are now discussed by reference to the experience of the United Kingdom.

5. RECORD SYSTEMS IN THE UNITED KINGDOM

Decennial censuses in the United Kingdom use conventional methods. The 1981 census was probably the most successful census since the Second World War — a success that was helped by the shortened form and the omission of a controversial question on ethnicity. So three factors combine to make a register-based census seem a rather remote possibility: the 1981 success; doubts about the range and quality of statistics that could be extracted from administrative records; and the absence of a population register to coordinate the record systems.

But statisticians have recognised the benefits, both administrative and statistical, that population registers could bring. The two initiatives on this subject in the past 70 years — both of which failed — are described in Sections 5.1 - 5.4. Now the government, while opposing a central population register, is introducing a limited form of local population register as part of a controversial reform of local taxation (Section 5.5).

5.1 National Registration in Two World Wars: The 1918 Committee on Registration

Thinking in Britain about population registers goes back over seventy years to the First World War. The National Registration Act of 1915 had obliged every adult to carry a National Registration Certificate and to register every change of address. This led Sir Bernard Mallet, Registrar General, to consider a permanent system, which he outlined in his Presidential address to the Royal Statistical Society in November 1916 (Mallet 1917). But he was aware that he might be criticised for "desiring to Prussianise our institutions".

These ideas were developed in the report of a committee appointed by the government in 1918 and chaired by Sir Bernard Mallet. Many years later he reviewed the findings in his Presidential address to the Eugenics Society (Mallet 1929). What he then said remains true today:

"We found in existence in England a very considerable number of registers being kept at considerable expense for various special purposes, some of them covering very large sections of the population. These registers are kept under different Acts of Parliament, by various authorities, in varying areas, for independent purposes, without any provision for their coordination one with another".

The committee proposed continuous registers of the population kept locally and associated with identity cards. A central index register would interrelate the local registers to deal with removals and to prevent duplicate entries. This registration system would coordinate the registers kept for special purposes — electoral registers, school attendance registers, the decennial census, registers of births, marriages and deaths, etc. It is noteworthy that the committee, reporting nearly seventy years ago, proposed that the census of population should be linked to population registration.

In his 1929 address Sir Bernard Mallet set out the principles to which any good system should conform: first, the accurate identification of every individual "in order (a) that he shall be made responsible for the fulfilment of his obligations to the community and (b) that he shall be ensured his rights as a citizen, whether these take the form of franchises to be exercised or dues to be received"; second, the acquisition of statistical information and in particular regular figures of the populations of local areas. The analysis made and the proposals that followed would still stand as a valid response to the situation that we face in the United Kingdom today, though some of the features would not be acceptable now. Thus:

"the numerous official enquiries and registers, now made and maintained independently of each other, would be coordinated into a single system which would provide a **dossier** for each individual containing those particulars regarding him which the State is concerned to know". (Mallet 1929)

To Sir Bernard Mallet's regret the recommendations in his committee's report were not carried out and, with the demise of the temporary wartime legislation, national registration ceased until the outbreak of the Second World War.

During the Second World War and for a few years after a full system of population registration operated in Britain. A National Register was set up linked to the issue to each person of an identity card bearing his identity number and address. Local registers were coordinated through a central register which held each person's name, date of birth, identity number and a code for area of residence. A person had to notify changes of address to the local register. The National Register survived until 1952 when identity cards and the obligation to notify changes of address were abandoned in a post-war spirit of "set the people free".

5.2 The National Health Service Central Register

The central register set up in 1939 during National Registration has been maintained since 1952 to serve a more limited role in the running of the National Health Service (NHS). Renamed the National Health Service Central Register (NHSCR), it now includes everyone resident in Britain apart from the 1 or 2 per cent who were born abroad **and** who have never registered on the patient list of a doctor in the NHS. But the NHSCR does not fill the role of a central population register of the kind found in many countries in Northern Europe because it is not used as a reference point from which other agencies can check personal identities and can carry the personal reference numbers into their own

files. Indeed the identity numbers recorded in the NHSCR serve only NHS purposes. Other limitations which would inhibit the wider use of the NHSCR are:

1) A significant proportion of the data arriving at the NHSCR do not carry the identity number and, given the difficulty in using names and dates of birth as unique identifiers, some of these data cannot be linked to already existing NHSCR records; thus some 1 or 2 per cent of the deaths notified to NHSCR cannot be linked in. This and the failure to remove all emigrants from the register are main factors in the inflation in the register, currently estimated at about 5 per cent. But this figure should reduce shortly when the register is computerised.

2) Addresses are held in full in local registers and as area codes in the NHSCR. But in most cases changes of address are recorded only when a person registers with a new doctor — which may occur years after the person has moved house.

5.3 The Wide Range of Registers in the United Kingdom

As in any other developed country, a wide range of registers containing personal data is held by public authorities in the United Kingdom. The main ones concern vital registration (births, deaths, marriages and divorces), immigration and naturalization, the national health service, social security (contributors and beneficiaries such as the unemployed, pensioners and children), personal taxation, passports, electoral lists, the ownership of cars and licences to drive cars. But these registers are maintained independently of one another by the different agencies, each with its own personal numbering system. (An exception is the joint arrangements for collecting employees' social security contributions and income tax under Pay-As-You-Earn, using one set of personal numbers, the National Insurance numbers.) This case apart, there is no coordination of record systems, no consistency in the content of records and no single set of personal numbers in general use. Details of a person's identity, usually name and date of birth, may differ between one register and another or even within the same register. This causes duplication and makes linking between registers for statistical purposes uncertain and costly. Information on address is even less consistent. There is no mechanism for carrying updating information simultaneously into all relevant records, for example information on change of address, change of name on marriage, or even the fact of death. In the words of Sir John Boreham, then head of the Government Statistical Service (GSS), "the information is never properly brought together ... It's all rather ramshackle" (Boreham 1985).

5.4 The 1960s Study of Registers

The existing uncoordinated system of records is inefficient for administration; and the absence of up-to-date addresses and the inability to link records are severe handicaps for statistics. And so in the late 1960s the GSS looked for a remedy. It studied the case for replacing the variety of personal numbering systems by a single set of personal numbers to be held in a central register, which might also include up-to-date addresses (Penrice et al. 1968). But Ministers decided that these ideas were politically unacceptable and terminated the studies (House of Lords 1969).

5.5 The Registers for the New Community Charge

It would seem that one of the biggest obstacles to the creation of a population register in Britain is now, in 1987, on the point of being overcome: an obligation is to be laid on the citizen to report changes of address. Despite this, no effective population register will be created. The government has set its face against that.

The new obligation to report changes of address — a revolutionary departure from peacetime traditions in Britain — stems from the government's decision to change the basis of local taxation. In the past local taxes have been levied on the occupiers of property on the basis of the property's rental value. The tax on the occupier of a dwelling is now to be replaced by a flat rate tax on each person aged 18 and over living in the dwelling: the **Community Charge** (CC). To administer the tax new local registers will be maintained listing addresses and the persons aged 18 and over resident there. Though the registration officer will be able to make enquiries and to call on information held by local authorities and housing bodies and in electoral rolls, the obligation to inform him of changes to the register is laid on the individual. Legislation has already been enacted to introduce the new system in Scotland with effect from 1989 (United Kingdom 1987), and the government intends to legislate for England and Wales in the present session of Parliament.

But the CC registers will be primitive instruments compared to the population registers in the Nordic and Benelux countries because:

1) The CC registers will not cover everyone; in particular they will not cover the under-18s and people living in boarding houses and institutions.

2) The registers (which will record each person's name, date of birth and address) will be maintained locally with a limited degree of standardisation of procedures. There will be no central register to standardise the description of each person's identity and to coordinate the local registers (for example to facilitate transfers between authorities).

3) Although the Scottish legislation makes no specific provision for including a personal reference number in the registers, a recent report has recommended that local authorities in Scotland should create such a number and suggested a possible algorithm for this based on name and date of birth (Chartered Institute of Public Finance and Accountancy 1987). But the Minister steering the legislation for England and Wales has said that, even if personal numbers are used in Scotland, they will not be needed in England and Wales (Howard 1987).

4) The Scottish legislation specifies who can have access to which parts of the register. Apart from local authority access for the purpose of administering the CC: an individual can inspect the entry relating to himself; the public can inspect the list of addresses and the names of persons relating to each address ("but not so as to ascertain whether that person resides at that address"); and the Electoral Registration Officer has access for his purposes. No other access is permitted.

The government's rejection of a population register that would coordinate administrative records is spelt out in the Green Paper on the CC scheme (Her Majesty's Government 1986). The paper cites countries that "have unified their separate registers and use them for several different central administrative purposes". It goes on "The British tradition is different. Registers are kept separately for different purposes by the body which needs them for a particular purpose. ... There will be no national register." This contrast between other countries' practices and United Kingdom practice is mistaken, because in other countries the different agencies maintain **separate** registers but call on a central register in order to identify the individuals that they are dealing with. I would judge that the statement "There will be no national register" reflects a political axiom, not the conclusion of rational analysis.

The creation of the CC registers is perhaps a missed opportunity to set up an effective population register. But the CC scheme is not an ideal vehicle for that. If it is to be effective, population registration should serve many ends, the more the better, and not just one — particularly when the single purpose is to levy a tax which many will feel

onerous and many may try to avoid. Moreover the CC is politically controversial because of its differential impact on various groups in the community: in general terms a transfer of resources from the poor to the rich.

Thus there are several reasons for questioning the operational effectiveness of the registers to be set up under the CC scheme: the single-purpose and controversial aim of the registers; the incomplete coverage of the population (the omission of some groups); the lack of a central register to coordinate the local registers; and the apparent reliance on a person's name and date of birth as identifiers rather than a permanent personal number. The local authorities have made some critical observations on the problems that they will face in attempting to set up the registers (Rating and Valuation Association 1987). It looks as though the government has embarked on new tax legislation without thinking through the practicalities of implementation.

Another worrying feature of the CC scheme is its effect on response to the 1991 census of population. Many of those who evade CC will probably try to evade the census too, not trusting the census authorities' assurances that census data will not be passed on to other agencies. And if the census form is too explicit by stating "YOUR INFORMATION WILL NOT BE PASSED TO THE AGENCIES DEALING WITH TAX, SOCIAL SECURITY, COMMUNITY CHARGES, ...", will the census authorities themselves be seen to be condoning or even encouraging evasion and fraud?

5.6 The United Kingdom Environment

Leaving aside the CC, the present environment in the United Kingdom is generally hostile to the idea of population registers. But two positive features may be mentioned. First, the Data Protection Act, 1984 introduced safeguards for personal data held on computers on the lines of the Council of Europe's Convention of 1981 (Council of Europe 1981). In fact the government's primary aim in introducing the 1984 legislation was commercial: to establish the United Kingdom as a safe place in the eyes of other countries which might be considering transmitting their data to the United Kingdom for processing. Protection of privacy was a lesser aim. Second the GSS, which would be concerned with some aspects of the working of population registers, has established an unquestioned record of protecting data; it has published a code of practice (Government Statistical Service 1984). Integrity in handling data has been underpinned by the fact that the GSS is decentralised, so that legal and administrative barriers have prevented the exchange of data even for statistical purposes. Such barriers would have to be removed if the statistical fruits of population registration were to be secured.

On the other side of the balance sheet the GSS's dependence on central government contrasts with the relative autonomy of the statistical organisations in, for example, Denmark and the Netherlands; this could lessen public confidence in its handling of data. The GSS's image as a creature of central government has been intensified by the Rayner Reviews of the early 1980s, as a result of which the GSS was instructed to give greater priority to the needs of central government at the expense of the needs of others — the local authorities, business, academics and the general public.

A main obstacle to population registers in the United Kingdom is the public's traditional resistance to governmental actions that appear to be overbearing or bureaucratic. The privacy lobby can be relied on to lead the opposition to any new reporting obligations placed on the public, to any extensions of the government's holding of personal data or to any project for linking data. The opposition overlooks the costs and injustices that result from inefficient management of data; and it overlooks or undervalues the checks on the misuse of personal data that can be provided by legislation on data protection and freedom of information — if properly implemented. In recent years fears about giving more personal data to the government have been reinforced by the public's perception of the style of government: the United Kingdom government is

seen as almost obsessively secret and as seeking to concentrate power in its own hands. Thus, not only is there no Freedom of Information legislation in the United Kingdom, but all government information is, in principle, protected by the catch-all Official Secrets Act, 1911. Peter Hennessy, writer and broadcaster, asserts that British governments "maintain the tightest system of administrative secrecy in the western world" (Hennessy 1987). And recent events have called into question the proper accountability of the security services. Writing of the whole range of government activity, William Plowden, Director General of the Royal Institute of Public Administration, said "a modern British government, supported by an adequate majority in the House of Commons, at little risk from the rubber-toothed bulldogs of the select committees and entrenched behind the Official Secrets Act, is one of the least accountable executives in the developed world" (Plowden 1987).

So the public is suspicious of any new scheme of population registration. And, as already noted, opposition to full registration has been expressed by the present administration, which, like its counterpart in the United States, has made determined efforts to "get government off our backs". One of the administration's major policy objectives has been to reduce the size and influence of the public sector — sometimes giving a higher priority to this than to cost-effectiveness. So public concern about privacy, political ideology and scarce resources combine to block a full register which could lead to substantial savings and to a fairer and more just society. In fact there has been no balanced presentation of all the issues, and so no public discussion of them, in the past half century.

6. AN AUSTRALIAN INITIATIVE: IDENTITY CARDS

I know little about the Australian temperament or the Australian political scene, but I guess that resistance to bureaucratic government is as strong there as it is in the United Kingdom. Even so, the Australian government introduced a Bill to issue each citizen with an identity card — the Australia Card (AC). The reasons were wholly administrative: to reduce tax evasion, to reduce social security fraud and to reduce illegal immigration. The AC would carry the person's name, his photograph, his signature and an AC number (personal reference number) but not address. It would be backed up by an AC register (which would also include address and date of birth) accessible only to certain government departments.

The Australia Card Bill, 1986 was passed by the House of Representatives but was rejected by the Senate (in which the government party did not have a majority). The rejection was given as one of the reasons for calling the July 1987 general election and, following the electoral success of the government party, the Bill was due to come before Parliament again. But the Bill has been withdrawn because of a serious legal flaw. However it is worth describing the Bill's provisions.

The AC register would be a central population register. But it would be less developed than those in Northern Europe for two main reasons:

1) The Bill did not place an obligation on the citizen to notify each change of address. The hope was, I understand, that most changes of address would be picked up by one or other of the government agencies taking part in the scheme and would then be passed on to the AC register.

2) The AC scheme would not be as multi-purpose as several of the population registers in Europe. As a result of concerns about privacy and uncontrolled linking of data, the AC register would be accessible only to the government agencies dealing with tax, social security and health insurance, and then only to check identities.

The Bill defined the situations in which a person could be required to produce his AC; these included making any of a wide range of financial transactions, entering a new employment, claiming Medicare or social security benefits, and receiving hospital treatment. It would be illegal to require a person to produce his AC in any other situation.

As a further safeguard on privacy the Bill provided for a Data Protection Agency. However the government argued that privacy had to be balanced against the losses to government funds through tax evasion and fraud. The government estimated that the costs to government of the AC scheme would $0.8 billion over ten years, but that this would be offset many times over by savings of $4.1 billion in tax and $1.4 billion in social security, giving a net saving over the ten years of $4.7 billion (Australian House of Representatives 1986).

Remarks made by the Minister of Health in Parliament (Australian House of Representative 1986) show what Ministers were trying to achieve and the clear political commitment:

> "I bring before Parliament today ... a long overdue reform to provide fairness and equity for all Australians."
>
> "No one doubts that the Australia Card will check tax evasion; no one doubts that it will contribute to the integrity of our social security system; no one doubts that it will be a useful weapon in deterring illegal immigration; no one doubts that by facilitating the pursuit of the money trail it will provide an invaluable instrument against corporate and organised crime."
>
> "Irrefutably, citizens need to be protected against abuse of their privacy by government. But equally citizens need to be protected against others who cynically hide behind the mantle of privacy to create false identities and thus defraud the community."
>
> "It is inevitable that this country will establish an identification system before the century is out."

Though the AC Bill has now been withdrawn, the government is searching for other ways to clamp down on tax and social security fraud, and so the story is not yet ended.

6.1 Identity Cards

The main emphasis in the Australian scheme was placed on the identity card as a way of checking identity, rather than on the personal number and register. Some European systems also combine the issue of identity cards with population registration; the Belgian system is one of the most highly developed. And undoubtedly the identity card provides an extra tier of security — provided it is not forged or stolen. In some countries identity cards are unconnected with population registration, for example in France.

In countries unaccustomed to identity cards in peacetime, the card is seen as a symbol of an authoritarian régime and an affront to civil liberties. That may be one of the reasons why the AC scheme generated so much public opposition in Australia. But much of the benefit from population registers can be secured without identity cards provided that citizens know their personal numbers and quote them in dealings with public authorities. This is what happens in Denmark and Sweden where population registration is effective, both administratively and statistically, without issuing identity cards to everyone.

A country like the United Kingdom ought not to shy away from correcting the incoherence of its records just because the uninformed critic might equate the necessary remedy — population registration — with what is only an optional extra — identity cards.

7. CONCLUDING REMARKS

Setting up a population register, with up-to-date addresses and personal reference numbers that are also carried into administrative files, would in fact be little more than bringing order into an existing "ramshackle" system: even in the most ramshackle system the citizen has to identify himself and inform various agencies of a change of address. Nonetheless some people are deeply worried by the prospect of a population register because of its threat to privacy and freedom and because it gives increased power to the State with all the dangers of misuse by an authoritarian or oppressive government. But specific remedies can and should be put in place: an effective data protection régime and legislation on freedom of information.

On the other hand a properly coordinated record system would have political advantages that have been largely overlooked. At the top of the list I would put two things:

1) A brake on fraud, crime and illegal immigration.

2) A fairer society, so that burdens and duties are fairly shared and benefits and rights go only to those entitled to them. Put another way, freedom should not extend to the freedom to cheat the rest of the community.

Rather lower down the list I would put:

3) The financial savings to government. More accurate records will cut the costs of administration, give a higher yield of tax and reduce the amount of benefits paid improperly — illustrated by the Australian figures (Section 6).

4) A wider range of policy options for government. Thus, if a reliable population register were already in place in the United Kingdom, the government would not have to construct a register **ad hoc** in order to launch its Community Charge scheme; and it could regulate immigration through control on residence in addition to the controls at airports and seaports.

5) Other benefits from more reliable checks on identity. The late Registrar General gives as an example better checks on a couple's eligibility to marry. There would also be fewer different reference numbers to be quoted and perhaps fewer plastic cards to be carried.

6) Better statistics (but see a qualification made following table 1).

This list is one answer to the charge that a population register is totalitarian and Big Brother. Without safeguards and in the wrong hands it could be. But it could also be the key to a fair and just society. The question is: what kind of society do we seek? Is it one that encourages, or at least turns a blind eye to, fraud, tax evasion and crime? Australian Ministers cite the man who was convicted for collecting over 50 separate unemployment benefit cheques each fortnight (Australian House of Representatives 1986). In the United Kingdom a Member of Parliament and barrister who made multiple applications for shares against the rules by using different names, addresses and bank accounts has just received a prison sentence (subject to appeal); the defence was that it was common practice.

Another answer to the charge of totalitarianism is to look at the population registers in other countries. Table 1 divides 15 countries — all the countries of Western Europe except Austria and Switzerland — into four groups according to the kind of register system that each has. The six countries in group A have the most effective systems: their administrative records are coordinated by the population registers. The four countries in group B are in an intermediate position. In the three countries in group C

population registers exist only at the local level and their quality is sometimes poor. Finally Ireland and the United Kingdom are in group D at the least developed end of the spectrum. If the United Kingdom were to take what I believe is a rational and realistic course and move into group A, it would not be joining a totalitarian company.

The statement noted earlier (item 6) that a properly coordinated record system will lead to better statistics needs to be qualified. Better statistics are indeed the **direct** consequence; a good example is regular and reliable population statistics for small areas. But if, as an **indirect** consequence, irresistable pressure builds up to replace a conventional census by a wholly register-based census, there are both benefits and penalties. Against the benefits of lower costs, a smaller burden on the public and a lesser risk of sabotage has to be set the probable deterioration in the range and quality of census results on economic topics, housing etc. Thus administrative records may increasingly fail to reflect the complexities and informalities of present-day life-styles which a conventional census could attempt to record — for example more part-time employment and self-employment, more second homes and looser family and household ties. It is here that Nordic experience (Section 3) is relevant.

Statisticians are not likely to underestimate the value of better statistics. But policy and administration — political considerations — carry a bigger weight in the arguments for and against population registers. The arguments need therefore to be debated by policy-makers, politicians and the public. In the United Kingdom a debate ought to take place on the wisdom — indeed the feasibility — of constructing the proposed single-purpose CC population register deliberately disconnected from other registers, rather than a multi-purpose population register with all the benefits that that could bring.

But I believe it right to bring the subject before statisticians for three reasons. First statisticians understand both the technical problems and the wider issues, and so can give a lead. Thus, in the United Kingdom both the earlier initiatives on population registers were taken in a statistical-cum-registration context (Section 5). Second, statistical agencies may be given responsibility for the key coordinating mechanisms, in particular the central population register, as INSEE has in France and SSB in Norway. Third, statisticians would benefit from more reliable data.

I hope therefore that statisticians will make their views known. Registers are very much a live issue, not least in such "under-developed" countries as the United Kingdom and Australia. Statisticians working in government service should reflect on the comment on professional ethics offered to the US Bureau of the Census; the words were written in a different context by the 1984 Panel on Decennial Census Methodology (Citro and Cohen 1985) but are very relevant here:

> "We recognise that the temper of the times is not conducive to the initiation of new programs, but we believe that statisticians have the responsibility to describe the facts and recommend the actions they believe are sensible."

Table 1

Particular Features of Population Registration in 15 Countries - Indicated by the Symbol x
For Fuller Details see Redfern 1987

	Local Population Registers	A Central Population Register which Coordinates Administrative Records	Personal Reference Numbers
A. With a Full System of Population Registration			
Belgium	x	x	x
Denmark	x	x	x
Finland	x	x	x
Luxembourg	x	x	x
Norway	x	x	x
Sweden	x	x	x
B. Intermediate Group			
France	.	x	x
Netherlands	x	.	x
Portugal	.	x	x
Spain	x	(x)	x
C. With Local Population registers only			
F R of Germany	x	.	.
Greece	x	.	.
Italy	x	.	.
D. Without Population Registers			
Ireland	.	.	.
United Kingdom	.	.	.
Number of Countries with the Feature	11	8+	10

ACKNOWLEDGEMENTS

For the information used in preparing this paper I am grateful to the statistical offices of Australia, Finland and Norway and, of course, of the countries which contributed to my report to the EEC. Responsibility for errors and shortcomings is mine.

REFERENCES

Australian House of Representatives (1986). The Honorable Neal Blewett MP in the Second Reading Debate on the Australia Card Bill, 1986.

Boreham, J. (1985). Quoted in "How Whitehall plays the Numbers Game", *The Times*, London, 30 July 1985.

Chartered Institute of Public Finance and Accountancy (1987). Preparation of a specification of user requirements for the system of community charge in Scotland. Unpublished manuscript, CIPFA Services, London.

Citro, C.F., and Cohen, M.L. (eds.). (1985). The Bicentennial Census: New Directions for Methodology in 1990. Washington, D.C.: National Academy Press.

Council of Europe (1981). The convention for the protection of individuals with regard to automatic processing of personal data.

Government Statistical Service (1984). The Government Statistical Service code of practice on the handling of data obtained from statistical inquiries. Cmnd 9270, Her Majesty's Stationary Office.

Heinonen, R., and Laihonen, A. (1987). Some new solutions and methods for census data production: Finnish experiences from the 1985 census. Paper represented at the ECE/CES Seminar on Computer-Related Aspects of Population and Housing Censuses, Belgrade.

Hennessy, P. (1987). *The Independent*, London, 1 April 1987.

Her Majesty's Government (1986). Paying for local government. Cmnd 9714, Her Majesty's Stationary Office.

House of Lords (1969). The Lord Chancellor, Lord Gardiner, in Hansard, 3 December 1969.

Howard, M. (1987). Michael Howard, Minister of State, Department of the Environment on BBC Radio 2, 30 September 1987.

Jensen, P. (1983). Towards a register-based statistical system -some Danish experience. *Statistical Journal of the United Nations Economic Commission for Europe*, 1, 341-365.

Johansen, S. (1987). Input dubious — output OK. Paper presented at the European Population Conference, Jyväskylä, Finland.

Johansson, S. (1987). Statistics based on administrative records as a substitute or a valid alternative to a population census. Paper presented at the meeting of the International Statistical Institute, in Tokyo.

Laihonen, A., and Myrskylä, P. (1987). Use of registers and administrative records in population censuses in Finland. Paper presented at the European Population Conference, Jyväskylä, Finland.

Mallet, B. (1917). The organization of registration in its bearing on vital statistics. *Journal of the Royal Statistical Society*, Part 1, 80, 1-24.

Mallet, B. (1929). Reform of vital statistics. *Eugenics Review*, 21, 87-94.

Penrice, G., Redfern, P., Evans, D., Whitehead, F.E., Bishop, H.E., and Rudoe, W. (1968). Discussion of the papers on social and medical statistics, *Journal of the Royal Statistical Society*, Ser. A, 131, 26-33.

Plowden, W. (1987). The battles of ideology that ill serve the public. In *The Independent*, London, 24 June 1987.

Rating and Valuation Association (1987). Community charge, poll tax: the facts. Unpublished manuscript, Rating and Valuation Association, London.

Redfern, P. (1987). A study of the future of the census of population: alternative approaches. Eurostat Theme 3 Series C, Office for Official Publications of the European Communities, Luxembourg.

United Kingdom (1987). Abolition of Domestic Rates etc. (Scotland) Act, 1987.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

PROTECTION OF TAXFILER DATA

H.J. LAGASSÉ[1]

ABSTRACT

The use of taxfiler data for purposes other than tax administration requires balancing competing policy objectives. The primary objective for National Revenue Taxation is to ensure the confidentiality of taxfiler data and its use solely as permitted by law. The legal framework and the obligations and requirements of National Revenue Taxation impose significant limitations on the use of this data for statistical purposes.

1. INTRODUCTION

The Canadian income tax system is based on the principle of "self-assessment" whereby individuals and businesses provide National Revenue Taxation with information on themselves and their financial affairs. A high level of voluntary compliance by taxfilers in providing that information is fundamental to the success of the system. Voluntary compliance in turn depends on taxfilers knowing that the personal and financial information they give to the Department will be kept confidential and will be used only for purposes allowed by law. Various measures — legislative, administrative, and technical — have been adopted to help ensure that this condition is met.

On the other hand, the pursuit of effectiveness and efficiency and the reduction of respondent burden on individuals and businesses, argue for communicating taxfiler data to certain government organizations under prescribed conditions. The Income Tax Act and certain other statutes permit the communication of taxfiler data by National Revenue Taxation in particular instances.

A balance must be established and maintained among these competing values and goals. Achieving that balance is a continuing challenge, not only for officials of National Revenue Taxation, but also for officials in other federal departments and agencies who may receive taxfiler data. These latter officials are also bound by the confidentiality provisions of the Income Tax Act, and they may only use and communicate taxfiler data solely as permitted by law. They are liable to the penalties set out in the Income Tax Act if they contravene the law.

Recognizing that the tax system does not exist mainly, or even secondarily, to provide taxfiler data to other departments, officials of both National Revenue Taxation and other departments must be aware of, and sensitive to the real costs and benefits of National Revenue Taxation releasing tax data for purposes other than income tax administration. Officials of National Revenue Taxation also need to satisfy themselves that other authorized users recognize and accept their obligations and responsibility for those costs — both the monetary costs and the more intangible, but nonetheless real costs of any

[1] H.J. Lagassé, Director General Systems, Revenue Canada Taxation, 875 Heron Road Rm. 5140 Ottawa, Ontario. K1A 0L8

weakening of the public's trust in the confidentiality of their personal and financial information.

In this paper I review the main legislative and policy provisions affecting the protection[2] of taxfiler data and its communication to certain federal departments and agencies for specified purposes **other than** income tax administration. I will address all such uses since the organizational issues and policy considerations are not confined to statistical uses. However, the latter uses do give rise to a few particular issues from a tax authority's point of view, and I will note those briefly.

2. THE LEGAL FRAMEWORK

In general, certain federal statutes enable particular federal departments or agencies to obtain and use identifiable taxfiler data for certain specified purposes. The purposes range from administration, enforcement or evaluation of certain statutes or government programs, to the compilation of statistics and tax policy analysis. For example, the Statistics Act enables Statistics Canada to access taxfiler data for statistical purposes, while the Income Tax Act permits officials of National Revenue Taxation to communicate tax data to officials of the Department of Finance for purposes of tax policy analysis. However, other statutes, including the Income Tax Act and the Access to Information Act, contain provisions which limit the communication and use of taxfiler data. In some instances both the Income Tax Act and other statutes must be consulted to determine whether a federal department or agency may receive and use taxfiler data for a particular purpose.

In this section I outline the main statutory provisions affecting the use and protection of taxfiler data. It is helpful to begin with the Income Tax Act, whose pertinent provisions I will try to summarize in plain language.

Section 241 of the Income Tax Act does several things. It first establishes the cornerstone of confidentiality by directing "officials and authorized persons" not to communicate or allow to be communicated information obtained by or on behalf of the Minister of National Revenue **except as otherwise permitted.** The Income Tax Act defines "official" broadly as any person employed or formerly employed or occupying a position of responsibility in the service of Her Majesty. "Authorized person" has a similar meaning in the context of administering the Income Tax Act.

Given this broad definition of "official" it would not be prudent to stop there since officials could conceivably go on communicating taxfiler data to other officials or authorized persons, and confidentiality could quickly erode. I suspect that the drafters of the Income Tax Act recognized this risk. In any case, the Income Tax Act includes provisions making it an offence for **anyone** to contravene the above-noted cornerstone of confidentiality or to further communicate information **for any purpose other than that for which the information was first provided.** In other words, if an official in Department A receives taxfiler data for a specific purpose permitted by law, that official may **not** communicate the data to anyone else except for that specific purpose. We call this key provision the principle of consistent use.

Having established the cornerstone of confidentiality and the principle of consistent use, Section 241 of the Income Tax Act permits "officials or authorized persons" to communicate taxfiler data in certain **exceptional** situations.

[2] "Protection" refers to privacy, confidentiality and security.

Examples of such situations where the Income Tax Act permits an official or authorized person to communicate taxfiler data include:

- to an official of the Department of Finance solely for the purposes of evaluating and formulating tax policy;
- to an official of our sister Department — National Revenue, Customs and Excise — solely for the purposes of administering or enforcing certain Acts administered by that Department; for example, the Customs Act and the Excise Act;
- to an official of Canada Employment and Immigration Commission or the Department of Employment and Immigration (CEIC) solely for the purposes of administering, evaluating or enforcing the Unemployment Insurance Act or a prescribed employment program.

Government statisticians will know that Section 241 of the Income Tax Act also includes a provision permitting the communication of certain taxfiler data for the purpose of enabling federal and provincial government organizations to obtain statistical data for research and analysis. The data in question are limited to the name, address, occupation or type of business of a taxpayer.

Section 241 also includes a more general provision permitting the communication of taxfiler data to any person "otherwise legally entitled thereto". This is a significant provision when it is read together with other statutes such as the Tax Rebate Discounting Act, the Auditor General Act and the Statistics Act. Briefly, these statutes enable officials of the organizations concerned to access taxfiler data for specified purposes of those organizations, thereby establishing the "otherwise legally entitled" provision to which Section 241 of the Income Tax Act refers.

As indicated, discussion of the permissible release of taxfiler data beyond Revenue Canada Taxation centers on Section 241 of the Income Tax Act and sometimes other statutes. Consequently it is easy to overlook two other relevant sections of the Income Tax Act. Section 230 requires the Minister of National Revenue to give the Chief Electoral Officer a report on the aggregate amounts contributed to each registered federal political party and to each candidate at a federal election or by-election. That report is a public record. Finally, Section 149.1 requires the Minister of National Revenue to make public the information contained in public information returns filed by registered charities. The Minister may make public certain other information on charities; for example, their names, locations and registration numbers.

The preceding outline suggests the range of policy goals being pursued in addition to the protection of taxfiler data. The range of competing goals encompasses the effective and efficient administration of certain social and economic programs, national security, the efficient collection of statistics, the formulation and evaluation of tax policy, and, with respect to charities and political contributions, information to the public. And there may be other policy goals I have overlooked.

Two other pieces of legislation form parts of the overall legal framework for the protection of taxfiler data, and should not be overlooked. They are the Access to Information Act and the Privacy Act. The Access to Information Act provides for a right of access by individuals to records under the control of a government institution, while the Privacy Act provides the right of access by individuals to personal information about themselves.

For National Revenue Taxation, disclosure under the access legislation is governed by the confidentiality provisions of Section 241 of the Income Tax Act. Unless disclosure is permitted by Section 241 of the Income Tax Act, the confidentiality of personal and financial information contained in income tax files is protected at all times. The

following are some illustrative examples of requests for taxfiler information, made under the access legislation, that have been denied.

- A list of companies that have received scientific research tax credits.
- A list of companies that offer a deferred profit sharing plan to their employees.
- A list of most popular names.
- A list of individuals who claimed more than one thousand dollars of interest income.
- The number of life insurance companies and the total value of their reassessments.

As a layman in these matters I think that the effect of the access and privacy legislation is to limit the release of taxfiler data and to reinforce the confidentiality provisions of Section 241 of the Income Tax Act. To put it differently, the access legislation does not undo the protection of taxfiler data provided under the Income Tax Act, while the privacy legislation seems to me to reinforce those provisions of the Income Tax Act which enable a person, or his or her agent, to access personal information about themselves held by the Department.

Finally, in concluding this overview of the legal framework conditioning the protection of taxfiler data, I should mention two other protections. In the case of those federal government departments which may receive taxfiler data for specified purposes, most (perhaps all) have legislation which includes provisions similar, in varying degrees, to the intent of Section 241 of the Income Tax Act. For example, the Statistics Act includes a secrecy provision and penalties for breaching it. Also by way of example, the Auditor General Act has a similar confidentiality provision. Finally, all public servants are bound by their oath of secrecy. Taken together with the confidentiality provision, consistent use principle and related penalty provisions of the Income Tax Act, and with the exceptions to the right of access under the Access to Information Act, these latter statutory provisions provide what, in my view, are important additional measures of protection for taxfiler data.

3. ORGANIZATIONAL ISSUES AND POLICIES

Security Environment

As I noted in my introduction, a high level of voluntary compliance by taxfilers is fundamental to the success of the Canadian self-assessment tax system. This in turn demands that taxfilers have reasonable assurance that only authorized persons will have access to this sensitive information and that those persons will only use the information for purposes permitted by law.

While, as I have noted, the law includes provisions for the protection of taxfiler data, there are limits to what the law itself can do. Organizational policies and practices are at least as important.

In National Revenue Taxation the responsibility to ensure that only authorized persons have access to taxfiler data has always been taken very seriously. And we expect and insist that authorized persons in federal government departments receiving tax data also take this responsibility seriously.

The objective of the Department's security policy is to protect information, employees and property and to be able to give taxfilers assurance that the information they provide

will be held in confidence and used only for purposes permitted by law. To that end the Department has long had security policies and procedures in place and has provided security awareness training and information to its employees.

In June 1986 the Treasury Board of Canada issued a government-wide security policy. Its purpose is to prescribe a security system for the Government of Canada that will effectively protect classified information and other assets sensitive to the national interest from unauthorized disclosure, destruction, removal, modification or interruption. The policy is also intended to prescribe safeguards for other valuable assets and sensitive information such a taxfiler data.

The security policy requires compliance with **physical** security standards developed by the Royal Canadian Mounted Police (R.C.M.P.); with **personnel** security standards developed by the Treasury Board Secretariat; and with standards for **information technology** security developed by the R.C.M.P. and the Communications Security Establishment. Briefly, these standards establish the minimum mandatory requirements for protecting government information and assets and for the security screening and checking of employee reliability.

In November 1986, National Revenue Taxation began to implement the job-related personnel screening provisions of the new security policy. This is intended to ensure that all persons in the Department, who have access to departmentally sensitive information, meet the requisite standards of reliability, trustworthiness, and loyalty.

The Department is presently completing a systematic review of its information holdings to establish the appropriate levels of protection they demand, and is examining the physical and information security standards with a view of implementing them by early 1988. In the interim, existing security measures are being enforced diligently within the Department, and improvements are being made where needed and practicable. A recent case in point is the action we have taken with Statistics Canada to further strengthen the protection of taxfiler data in transit between our organizations in Ottawa.

As departments which may receive taxfiler data implement the government's new security policy and standards, the overall environment for the protection of taxfiler data should be enhanced, which we in National Revenue Taxation of course welcome. As well, the application of common security standards by those departments should help to simplify our common task of protecting taxfiler data.

Just as the provisions of the law can only go so far in providing protection of taxfiler data, security policy and procedures also have their limits. Other policies and practices exist within National Revenue Taxation to help assure both the confidentiality of information and its use only for purposes allowed by law. Increasingly those policies and practices are involving other departments.

4. CONTROL OF THE COMMUNICATION OF TAXFILER DATA

National Revenue Taxation has traditionally used several means to control the communication of taxfiler data.

Authority to introduce amendments to the Income Tax Act lies with the Minister of Finance. However, the obligations of National Revenue Taxation as the taxing authority give it a strong interest in the administrative implications of proposed amendments. Accordingly, proposed amendments to the Income Tax Act affecting the communication and protection of taxfiler data typically involve officials of National Revenue Taxation. The same is true of proposed amendments to other statutes which would affect the communication or protection of taxfiler data.

I would err if I gave the impression that we view any such proposals lightly, or if I appeared to encourage colleagues in other departments to think that such proposals can be easily incorporated into law. Indeed, the contrary is true. Proposals to amend the law affecting the protection of taxfiler data are examined with care by officials of National Revenue Taxation. They are also scrutinized by Members of Parliament. We in National Revenue Taxation are concerned with fully establishing the overall costs and benefits — to government, to the community and to the Department — of any such proposals. As well, we are concerned with ascertaining implications of such proposals in terms of National Revenue Taxation's operational priorities and requirements. And we advise our Minister accordingly.

Requests for data from departments having a defined entitlement (and others) have long been scrutinized. The first step has been to confirm entitlement to the particular data sought and the permissibility of the specific purposes for which it will be used. The second step has been to determine the availability of the data, the feasibility and the cost of producing it, and the time frame within which the request could reasonably be met, always bearing in mind National Revenue Taxation's obligations and requirements as the taxing authority.

Departments have been denied taxfiler data to which they had no entitlement in law. They have been denied data intended to be used for a purpose not permitted in law. They have withdrawn or modified requests because data could not be provided as or when requested, or at an acceptable cost. In other instances, departments have agreed to accept aggregate data in lieu of confidential taxfiler data when the former would in fact suffice for their purposes.

Despite potential economies or efficiency gains, National Revenue Taxation has not permitted linking its computer data bases to terminals or computers in other departments entitled to receive or use taxfiler data. Nor is any such linkage contemplated.

Care has also been taken to avoid releasing any more data than is necessary for a particular permissible use.

National Revenue Taxation recently reviewed and enunciated its internal policy on the release of taxfiler data for purposes other than tax administration. In keeping with that policy we are actively entering into written agreements with those federal departments with an entitlement to receive taxfiler data. Our aim is to better manage the release of data to them and to foster enhancement of the protection they accord to that data. We believe these agreements will prove to be of considerable value to both parties, not just National Revenue Taxation.

In general, these agreements identify the data that will be provided and the conditions and procedures that will apply to its release and protection. The agreements establish clear channels of communication between National Revenue Taxation and other departments with respect to the release of data, modifications to the agreements (to keep them current), and security matters.

Also in keeping with National Revenue Taxation's updated policy, our Systems Directorate is auditing its computer generated outputs to ensure that information released is the same as that authorized for release. As well, the Department maintains a record of all data released to particular departments.

5. STATISTICAL USES

To this point I have discussed the protection of taxfiler data from the perspective of National Revenue Taxation as the source of that data for permissible uses by certain federal departments. The Department also generates tax statistics or aggregated data,

some of which are published, for numerous clients both inside and outside National Revenue Taxation. In that context too the Department endeavours to assure the protection of taxfiler data, and to get its clients to accept responsibility for their use of the resulting data and the incremental costs of producing it.

In all cases where the provision of statistics or statistical services to a client require 10 or more person-days of work, we draw up detailed project terms of reference for approval by the client. In general, these terms of reference detail the statistics or services to be provided, the purpose and objectives involved, and the associated methodology, time frames, project milestones and required resources. In signing the terms of reference clients accept responsibility for both the products and their subsequent interpretation and use of the statistics.

Where statistics are produced for clients and other users not entitled to receive identifiable taxfiler data, standard statistical procedures are applied in order to effectively eliminate any potential residual disclosure of confidential information. Additionally, our major annual publication, **Taxation Statistics,** is periodically tested by our statisticians to isolate areas where changes may be warranted notwithstanding the care taken in designing the publication. In a field as dynamic as tax policy and its administration, such additional care is, I believe, more than merely desirable.

In closing I want to address two issues in the statistical uses of taxfiler data. These issues are of concern chiefly to others outside National Revenue Taxation, but they also affect us.

The first of these issues concerns the definition of terms such as "income" and "spouse". Terms such as these have particular meanings within the context of tax law and its administration. Those meanings can change with time and may or may not be entirely suitable for different organizations involved in compiling or using statistics, in the areas concerned, based on taxfiler data.

The second and related issue concerns the continuity of particular taxfiler data elements and the related implications for the production of consistent statistical time series. As a result of changes in the tax law and consequential changes in tax returns and schedules, information submitted by taxfilers can and does change (sometimes significantly) from year to year and even within a tax year. For example, the employment expense deduction has had a maximum value of $500 for several years. For the 1982 tax year the deduction was 3 percent of employment income up to the maximum. For the 1983 tax year and beyond the deduction has been 20 percent. An initial analysis of these data might well reveal blips resulting from the tax policy change but not underlying socio-economic conditions. Another example is the use of the taxfiler's address used to code geographic identifiers. National Revenue Taxation's interest is to have an address for contacting the taxfiler, mailing a refund or one of several other tax related reasons. Serious attention is not given as to whether the address is the principal residence, place of business, post office box or a neighbour's mailbox. However, for a statistical time series analyzing sub-provincial trends having to do with relative change between geographic areas, such differences can be of consequence.

In other words, there are significant limitations inherent in the use of taxfiler data for statistical purposes. My experience suggests that, by and large, persons using taxfiler data for statistical purposes understand and accept those limitations.

6. CLOSING OBSERVATIONS

Public policy and administration must usually seek to balance a number of competing goals. As I have tried to show, the use of taxfiler data for statistical and other non-tax purposes is a case in point.

National Revenue Taxation strives to fulfill its obligations as the tax authority effectively and efficiently. As well, it strives to fulfill these obligations in a manner that respects the rights of taxpayers, including their right to expect that the personal and financial information they provide will be held in confidence and used only for purposes allowed by law.

As I have noted, those purposes are numerous and varied. Officials of other departments involved in the use of taxfiler data are also bound by the confidentiality and consistent use provisions of Section 241 of the Income Tax. And we are all bound by the government's security policy.

The income tax system does not exist to provide data for purposes other than tax administration. Thus, trade offs and compromises are inherent in using taxfiler data for statistical purposes.

I believe the record shows that collectively we have done and are doing a responsible and credible job in balancing the numerous competing goals we must aim to achieve. Future success will continue to call on our resources of understanding, diligence and cooperation, which I am confident will be forthcoming.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

STATISTICAL USES OF ADMINISTRATIVE RECORDS IN THE UNITED STATES: WHERE ARE WE AND WHERE ARE WE GOING?

THOMAS B. JABINE and FRITZ SCHEUREN[1]

ABSTRACT

Given available resources and technology, the U.S. Federal statistical system can, if it chooses, place much greater reliance on administrative records. Important factors in shaping a strategy for future development of statistical use of administrative records include: the views of the statistical agencies, the custodians of the administrative records and the general public; the need to deal with frequent changes in the coverage, content and structure of administrative record sources; legal and policy provisions covering confidentiality, disclosure and access to administrative records; and the need for coordination among the statistical agencies in our decentralized system. Recent U.S. trends and changes related to each of the factors and their potential effects on administrative record research are reviewed.

INTRODUCTION

"Statistics" in its original meaning comes from the word "State" and meant "Facts about the State." Administrative records were the major source of such facts. With the advent of regular census-taking, modern societies began to have other good sources of data. Still, since the beginning, there have been people saying, "We have all these administrative records (about something or other). Surely we can get more out of them."

By and large such people have been right, but, as Tom Peters would say, "The obvious isn't always so obvious" (Peters and Waterman, 1982) -- certainly not to custodians of such records who typically operate on a tight budget and generally use the records for different purposes. The statistician soon learns, if he doesn't know it already, that doing the obvious, **provided he can convince enough people that it is obvious,** turns out almost never to be easy.

Once you decide to have a Conference like this, of people **who do what is obvious but not easy,** you feel (at least we do) compelled to define the field. Now, we are having trouble putting into words what those of us who are interested in statistical uses of administrative records **really are interested in.** As Gordon Brackstone (1987a, 1987b) has just said, the data sources we use span virtually the entire breadth of modern life. The techniques we employ also cover much of modern statistics, although there are some core technologies that we tend to focus attention on, e.g., record linkage (Internal Revenue

[1] Thomas B. Jabine, Statistical Consultant, 3231 Worthington Street, N.W., Washington, D.C. 20015, USA;
Fritz Scheuren, Director, Statistics of Income Division, Internal Revenue Service TR:S, 1111 Constitution Avenue, N.W., Washington, D.C. 20224, USA.

Service, 1985; Howe and Spasoff, 1986) and synthetic estimation (e.g., Gonzalez and Hoza, 1978; Fay and Herriot, 1979; Hidiroglou et al., 1984).

Maybe some progress can be made by looking at the kind of people who work in this field. Martin Wilk (1985), a few years ago when he was the head of Statistics Canada, wrote a paper about the differences between "blue-collar" and "white-collar" statisticians. Many of you here may remember his definitions. "Blue-collar" statisticians were basically the compilers of facts; "white-collar" statisticians were essentially academicians, or those interested primarily in theory.

The field of statistical uses of administrative records has necessarily and rightly been the province of "blue-collar" types, mostly in government. We don't want to overdraw this, but the emphasis has been on refining technologies, not inventing new subdisciplines of statistical science. A good example of what we mean is the work of Howard Newcombe (Newcombe et al., 1959; Newcombe, 1967) who really developed the technology of record linkage with his colleagues well before the formal Neyman-Pearson framework of hypothesis testing was set up for the problem by Ivan Fellegi and Alan Sunter (1969).

Maybe you'll agree with us that the statistical uses of administrative records is dominated by "blue-collar" statisticians. This is not to say there isn't a role for "white-collar" types. In fact, as might be expected, this Conference features more "white-collar" work than does the field generally.

The toughest part of this area, of course, is that we don't just have "white-collar" and "blue-collar" statisticians; we have other professions involved, too. Because what we do really matters to people and to our societies, there are policy-makers, ethicists and still others concerned about the possible misuses of what we do (e.g., Flaherty, 1979; Cox and Boruch, 1985; Gastwirth, 1986). We would dub these individuals "clerical-collar" types, except that you might feel we were being too glib about a very fundamental part of this field.

Our presentation is about what we have been doing in the U.S. in the field of statistical uses of administrative records. This topic is obviously hopelessly broad. Also, our observation of the Canadian work in this area leads us to believe that while you have barriers of a different character from ours, you are quite a bit ahead of us in both the white and blue (and maybe even the clerical) collar parts of this field. Even so, maybe you would find it of interest to review, from a U.S. perspective, some of the same ground that Gordon Brackstone just did from a Canadian perspective.

The context for our remarks, comes from a comprehensive review of U.S. statistical uses of administrative records, which we undertook about four years ago. At that time, we proposed six goals for the field for the next 10 years and a strategy for achieving them (Internal Revenue Service, 1984; Jabine and Scheuren, 1985). John Leyes of Statistics Canada chaired the American Statistical Association panel session at which we gave a paper on this topic. The other participants represented several U.S. statistical agencies (Butz, 1985a; Carroll, 1985; Norwood, 1985; Waite, 1985).

This Conference on the Statistical Uses of Administrative Data is an appropriate occasion to review the progress that has been made toward reaching these six goals, and we do that in the first section of our paper. To anticipate the findings of that review, progress has been less than we hoped for four years ago. Why is this so? Are the strategies we suggested inadequate? Are the obstacles to increased use of administrative records more formidable than we realized?

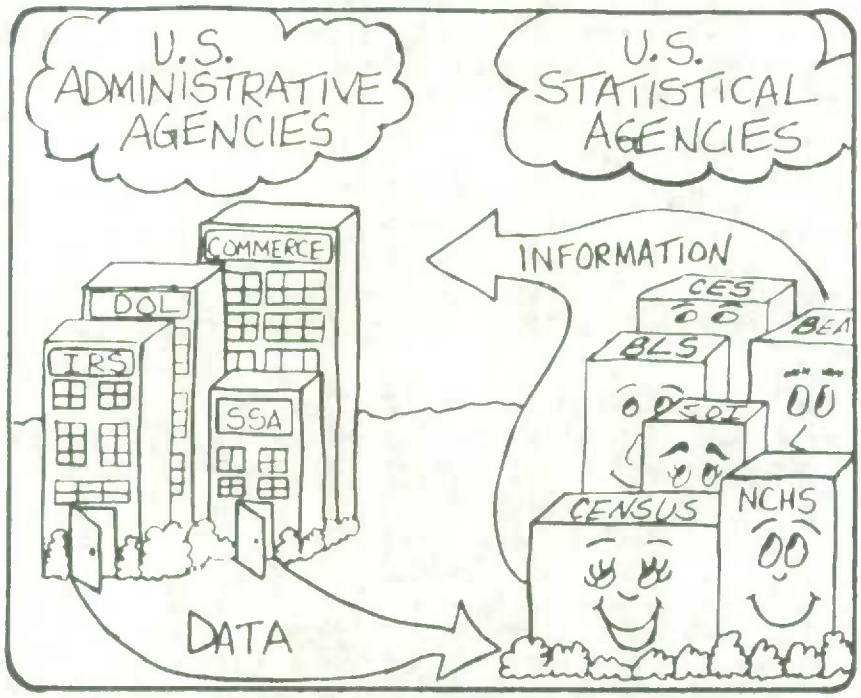
In looking for answers to these questions, we have identified a few of the elements which we believe to be the primary determinants of the extent to which administrative data can be exploited to their full potential for statistical purposes. These elements are reviewed in the second section of the paper, with emphasis on developments in the four

years since the six goals were proposed. This is followed in the third and final section of the paper with a new forecast for how we will (or might) deal with the future.

1. PROGRESS IN MEETING THE SIX GOALS

Before talking about the specific goals that we had the hubris to propose, we'll briefly discuss the differences and similarities between the U.S. and Canadian ways of organizing to do government statistics. The basic structure is the same in both countries. (See Figure 1.) There are statistical agencies and administrative departments. While statistical agencies have many sources of data, including data they collect directly, one of their chief sources is data from administrative agencies. Figure 1 shows just a few administrative agencies, each of which has a Canadian counterpart.

Figure 1



NOTE: Administrative Agencies

DOL = Department of Labor;
SSA = Social Security Administration;
IRS = Internal Revenue Service.

Statistical Agencies

CES = Center for Education Statistics;
BEA = Bureau of Labor Statistics;
SOI = Statistics of Income Division, IRS;
NCHS = National Center for Health Statistics.

It is on the statistical agency side that differences exist. In the U.S. there are several separate statistical agencies (NCHS, Census, BEA, etc.) that here would be centralized in Statistics Canada. But the Canadian system is not completely centralized either, since in the administrative departments in Canada there generally are statistical units, too. One of the ones we're most familiar with is the counterpart of the Statistics of Income (SOI)

Division at IRS (the one with the big eyebrows) — the Statistical Services Division at Revenue Canada Taxation. The Statistics of Income Division at IRS and the Statistical Services Division at Revenue Canada have very similar missions, too (as can be seen in part by contrasting two of this conference's papers on corporate income tax return statistics, by Fred Hostetter, Chris McCann and Brigitte Zirger of Revenue Canada (1987) and Hinkins, Jones and Scheuren (1987) from the Internal Revenue Service. (See also the paper by John Czajka (1987)).

As in Canada, administrative records are a major element in the Federal statistical programs of the United States. The three broad areas of uses (illustrated in Figure 2) are:

— **Program statistics**

This is a well-developed field and we had no goals here explicitly, although we regret not talking more about the need for improved policy simulation modelling (Revenue Canada, Taxation, 1985). The recent landmark Tax Reform Act of 1986 pushed our capacities to the limit in this area and pointed (again) to weaknesses that need addressing (Bristol, 1985). The paper by Wolfson et al. (1987) at this Conference fits in this broad area.

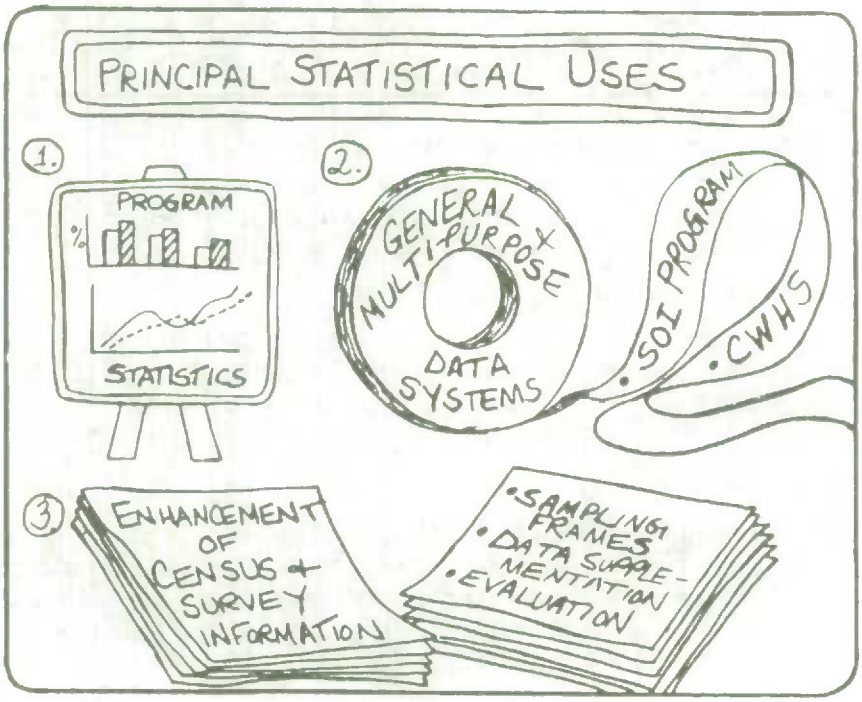
— **General purpose systems**

By general purpose systems we mean statistical data based on administrative records for which general (not just program-specific) purposes are served. Examples would include Social Security's Continuous Work History Sample and a lot of what is done with tax records as part of the Statistics of Income program. As you will see, we had two goals in this area (Wilson, 1983; Buckler and Smith, 1980; Kilss, Scheuren and Buckler, 1980).

— **Enhancing statistical data collection**

In this category are uses of administrative records as a frame, to augment survey data, etc. Most of our goals and, indeed, most of this Conference focuses on this area of use.

Figure 2



Six Goals

When we formulated the six goals for expanded statistical uses of administrative record (shown in Figure 3), we believed (and still do) that the use of administrative records in statistical programs is likely to increase further, given the ever-increasing costs of direct data collection and the general consensus on the need for reduction of the overall burden on the public to report information for both administrative and statistical purposes.

We felt, therefore, that U.S. statistical agencies and the statistical system as a whole should adopt a coherent, activist strategy for developing new statistical uses of administrative records. We saw a need for system-wide planning to make the best possible use of administrative record systems and to seek some degree of control over features of those systems that affect their suitability for statistical applications. The objective of the six goals was to focus attention on particular applications for which it seemed to us that the potential payoffs, for both statistical agencies and data users, were greatest.

Figure 3

Goals for Statistical Use of Administrative Records: The Next Ten Years (1984-94)

Goal 1: Explore fully and develop the uses of major administrative record systems in conducting and evaluating the decennial population censuses and in making current population estimates.

Goal 2: Starting with the Continuous Work History Sample (CWHS) as a base, develop a Linked Administrative Statistical Sample (LASS), using administrative records from several agency sources to construct a longitudinal person-based system that will serve a broad spectrum of users.

Goal 3: Give high priority to statistical applications of administrative records that will enhance our ability to monitor and analyze long-term environmental health effects.

Goal 4: Make greater use of administrative records in all phases of household surveys.

Goal 5: Develop and activate a "business directory" for use as a sample frame for economic censuses and surveys and a source of geographic and industry codes for common use by all eligible Federal and state statistical units.

Goal 6: Continue efforts to develop more consistent and compatible procedures for defining and identifying reporting units in both administrative and statistical data collections.

Source: Internal Revenue Service, 1984.

In this section, we review the six goals one by one. For each goal, we first describe the status of relevant activities as of mid-1984. Then, we review subsequent developments and provide our subjective evaluation of how far progress toward meeting that goal has succeeded in meeting expectations.

1.1 Use of administrative records for decennial censuses and current population estimates

Goal 1: Explore fully and develop the uses of major administrative record systems in conducting and evaluating the decennial population censuses and in making current population estimates.

Status in 1984

In 1982, Alvey and Scheuren proposed that a population census based on administrative records, primarily from IRS and SSA, be given serious consideration as an alternative to the traditional direct enumeration. The Census Bureau rejected this proposal for the 1990 census; however, it was given careful consideration and the internal working group that evaluated it recommended that a large-scale test of administrative record census procedures be undertaken as part of the 1990 census (Bureau of the Census, 1983). Childers and Hogan (1984) reported the results of a methodological study which linked a sample of tax filers aged 18 to 64 to the 1980 census. An important goal of the study was to investigate the feasibility of using the IRS Individual Master File as a frame for matching to the census to estimate gross undercoverage in the census. They concluded that this approach to coverage evaluation was promising and said that "the Bureau intends to continue research into the use of nonhousehold sources for coverage evaluation." However, the Census Bureau decided not to use IRS records in connection with pretests to be conducted in 1985 (but see **Current Status,** below, for subsequent uses).

The interim report of a Panel on Decennial Census Methodology, established by the Committee on National Statistics for the Census Bureau (National Research Council, 1984), did not include definitive recommendations on all uses of administrative records for the 1990 census. It recommended against immediate field testing of multi-list or composite-list methods for census estimation, adjustment or evaluation. Uses of administrative records for content evaluation studies and possibly for the collection of certain kinds of housing data were, however, encouraged.

Current Status

The role that will be played by administrative records in the 1990 census may be heavily influenced by final decisions on the controversial issue of adjustment of the census counts. The Commerce Department announced on October 30 that it "does not intend to adjust the 1990 decennial population count for purported undercount and overcount of population subgroups" (**New York Times,** 1987). However, a bill has been introduced in the U.S. Congress that would require adjustments (**Wall Street Journal,** 1987). If the Census Bureau is required to make official adjustments at the local level, data from major administrative record systems may provide significant inputs to the adjustment process. On the other hand, if the idea of adjustment is rejected, it seems likely that there will be less use of administrative records for coverage improvement in 1990 than there was in the 1980 census. Whatever the outcome on coverage **adjustment,** new uses of IRS records for coverage **evaluation** are being considered. The most likely possibility is the use of taxpayer name and address lists to check the completeness of field address listings for an independent coverage evaluation survey. The Census Bureau plans to include an administrative list component in the post-enumeration survey for the 1988 Decennial Census Dress Rehearsal in St. Louis, Missouri. Administrative lists to be used include: IRS Individual Returns Extract with matched SSA age, sex and race information; State Driver's License files, Draft Registration files; Veterans Administration files; and Unemployment Insurance files. In any case, it is possible that a limited study, at least, of administrative record census taking will be part of the Census Bureau's 1990 Research, Evaluation, and Experimental (REX) program.

Although an administrative record census is clearly out of the question for 1990, recent developments may make this approach a stronger contender for the year 2000. The Tax Reform Act of 1986 requires that social security numbers (SSNs) be reported for all taxpayers and for all dependents aged 5 and over, so that basic age, sex and race information will soon be available from the IRS/SSA records systems for a much greater proportion of the total population.

This development also has significant implications for current population estimates. The General Revenue Sharing program, which required current population and income estimates for some 39,000 local government units, has been phased out; nevertheless, IRS/SSA data continue to play an important role in the Census Bureau's small-area population and per capita income estimates program. Inclusion of dependents' SSNs in the IRS extract files provided to the Census Bureau would substantially increase the proportion of the total population for which direct estimates of internal migration could be made by linking IRS extracts for different years.

In a separate undertaking, IRS and SSA records have been used to produce 1985 data on population by ZIP (postal delivery) code areas. Estimates are based on Individual Master File extracts linked with SSA records that provide age, race and sex for a 20-recent sample of all persons with SSNs. Data on population by Hispanic origin were obtained by Spanish surname coding of the entire IRS file (as described in Passel and Word, 1980, 1987).

1.2 Enhancement of the Continuous Work History Sample (CWHS)

Goal 2: Starting with the Continuous Work History Sample (CWHS) as a base, develop a Linked Administrative Statistical Sample (LASS), using administrative records from several agency sources to construct a longitudinal person-based system that will serve a broad spectrum of users.

Status in 1984

The CWHS, which is maintained by the Social Security Administration (SSA), is a system of statistical files containing demographic and longitudinal earnings data for samples of persons who have been issued SSNs. Through the mid-1970s, CWHS microdata files were available to researchers in and out of government and were widely used for analytical research on labor market behavior, lifetime earnings patterns, internal migration, regional workforce characteristics and many other topics (U.S. Department of Health, Education and Welfare, 1978). Plans had been developed to enhance the CWHS system by links with data from other systems covering occupation, current residence, income taxes, social security benefits and mortality (Kilss, Scheuren, and Buckler, 1980). However, both the plans for enhancement and the general release of microdata files were brought to a rather abrupt end by the Tax Reform Act of 1976, which placed new restrictions on the statistical uses of employer and individual earnings data initially collected by IRS, and used as inputs to the CWHS (Buckler and Smith, 1980; Duleep, 1986). In 1984, negotiations were underway with the Bureau of Economic Analysis (BEA) for the resumption of limited releases of current CWHS files to that agency, which had been an important user and had also played a significant role in processing microdata files for distribution to other users.

Current Status

Efforts to release CWHS files to BEA were thwarted by potential disclosure problems and are no longer being actively considered. With support from the National Cancer Institute (NCI), IRS and SSA have undertaken a pilot project to create files that would link

CWHS records with IRS occupation and industry data, and with mortality and occupation information obtained from death certificates identified as a result of searches against the National Death Index (Crabbe, Sailer and Kilss, 1983). Death certificate data on cause of death and occupation have already been linked to a subset of the individual SOI sample, deliberately designed to overlap with the CWHS sample. The addition of several years of CWHS earnings data to the file is planned for the latter part of 1988. A key requirement of the project is the development of interagency agreements that will permit release of a public-use version of the linked microdata file to NCI and other users. Despite the complexity of such agreements, considerable progress has been made.

It is now 10 years since the suspension of general release of CWHS microdata files. Even though there have been limited achievements in the NCI pilot study, it is hard to be optimistic about a return to the access policies in force prior to the 1976 Tax Reform Act, let alone further enhancement of the system. SSA, although it continues to maintain the CWHS for internal use (e.g., Kestenbaum, 1985, 1986; Kestenbaum and Diez, 1981), is not in a position to support the improvements and development of access mechanisms that would be required to satisfy the needs of a broad group of users. It had been hoped that release of current CWHS files to BEA would allow that agency once more to evaluate the quality of industry and geographic codes developed under SSA's establishment reporting program. Without such an evaluation, it is impossible to determine whether the quality of these elements of the system has deteriorated subsequent to the introduction of annual wage reporting in 1978, as some people suspect. (The ERUMS project, described in Section 1.6 below, is, however, providing some of the facts needed to conduct such an evaluation.)

Perhaps we must reluctantly abandon visions of returning to the "good old days" and look at other possibilities for multi-purpose longitudinal data systems based on administrative records. One vehicle that might be able to meet some, but not all, of the same requirements is the IRS individual Statistics of Income (SOI) sample, for which a major redesign effort is in progress. The redesign will strengthen the longitudinal component in these annual samples and, thanks to the new requirements for reporting dependents' SSNs, will provide data for family-like taxpayer units, thus going a step beyond the CWHS, which is strictly a sample of persons (Jabine, 1987b; Hostetter, 1987). Any optimism about these developments must, despite a twenty-five year tradition of public-use files, be tempered by the knowledge that many of the same disclosure limitations will affect access to this data system (Strudler, Oh and Scheuren, 1986); also, it is clear that IRS by itself cannot be expected to provide the additional resources needed to enhance the system and make it more useful and accessible to a diverse population of users. Finally, some form of general access to **limited** extracts from the CWHS retains considerable promise as a feasible short-run goal.

1.3 Enhance ability to track long-term environmental health effects

Goal 3: Give high priority to statistical applications of administrative records that will enhance our ability to monitor and analyze long-term environmental health effects.

Status in 1984

The National Death Index (NDI) had become operational in 1981, with information for all registered deaths from 1979 onwards, and had already been used in several health and medical research projects (Patterson and Bilgrad, 1985). In our discussion of this goal, we urged the linkage of NDI mortality data to a large-scale statistical database such as the CWHS. We also pointed to ongoing work on occupation coding on death certificates (Crouse et al., 1983), and linkages of Current Population Survey (CPS) samples to mortality data via the NDI (Rogot et al., 1983).

Current Status

As discussed in subsection 1.2, the utility of linkages of mortality data to the full CWHS depends on the ongoing pilot efforts involving SSA, IRS, and NCI. In addition, linkage to the planned SOI longitudinal sample, which includes occupation data, may be feasible.

Other relevant activities show progress. As the result of a joint project by NCHS, the National Cancer Institute and the National Institute for Occupational Safety and Health for occupation and industry coding on death certificates, 32 states and the District of Columbia are now coding these items, and coded data for 23 of the areas is being incorporated by NCHS into its mortality data base.

The CPS linkages to the NDI are proceeding and results of the first round, covering deaths for 1979 to 1981, will be published shortly. Linkages for the years 1982 to 1985 have already been performed, and additional linkages are planned for 1989 to 1991 (Johnson et al., 1985; Rogot et al., 1985).

The NDI is an invaluable resource for studying the effects of environmental exposures on mortality. For study cohorts with reasonably accurate identification information, deaths to individual cohort members since 1979 can be determined reliably and inexpensively. However, if cause of death, occupation or other statistical information is needed, the situation is less satisfactory: most researchers must go to the vital registration offices in each of the states where deaths have occurred. For national studies, this process can be costly and time consuming.

1.4 Greater use of administrative records in household surveys

Goal 4: Make greater use of administrative records in all phases of household surveys.

Status in 1984

Exact matching techniques had already been used in many instances to link data from administrative records to data collected directly from individuals or households in surveys. In addition to such direct enhancement of survey data, there are many other ways in which administrative records have been or could be used in surveys, e.g., in stratification of sample units, development of frames, imputation of missing data, estimation and evaluation. An obstacle that had been encountered to some proposed linkages of survey and administrative data was the "reidentification" problem, i.e., the possibility that custodians of administrative files might be able, by matching against their own records, to reidentify individuals whose linked survey and administrative records were included in a public-use microdata file.

Current Status

Direct linkages of survey and administrative data continue. Of particular interest is the NCHS' recent policy of collecting SSNs and other identifiers in current surveys, such as the National Health Interview Survey, for use in subsequent followup studies that require linkage to administrative records, such as death certificates and the Medicare files (Scheuren, 1985). This policy has added a new dimension to the NCHS data — their use for prospective epidemiological analyses as opposed to just cross-sectional descriptions of health status, health care utilization and health-related behavior.

The reidentification problem has not been resolved, however, and the Census Bureau now is much less likely than in the past to release public-use microdata files which combine administrative data with data from surveys conducted under the authority of Title 13. It would not be possible today to duplicate earlier releases of files such as those

produced from the 1973 Exact Match Study or the Longitudinal Retirement History Survey. As a consequence, any survey sponsor who wishes to link survey and administrative data and have access to the resulting records cannot take advantage of the decennial census or the Current Population Survey as a sampling frame.

Other uses of administrative records in household surveys continue, especially for evaluation purposes. For example, two papers presented in this symposium (Moore and Marquis, 1987; Bowie and Kasprzyk, 1987) describe the Census Bureau's uses of administrative records in SIPP and other Census Bureau surveys. Tippett (1987) described the use of records of tax assessors and utility companies to evaluate data collected in the American Housing Survey and pretests for the 1990 Census. The Energy Information Administration continues to use records of utility companies and fuel suppliers to supplement data collected from households in its Residential Energy Consumption Survey (Energy Information Administration, 1987). There are still many additional opportunities for using administrative records in household surveys. All in all, the increased level of activity in this area is encouraging.

1.5 Development of a shared "business directory"

Goal 5: Develop and activate a "business directory" for use as a sample frame for economic censuses and surveys and a source of geographic and industry codes for common use by all eligible Federal and state statistical units.

Status in 1984

In November 1983, the Administration had withdrawn its support from a proposal for broad confidentiality legislation — the so-called "Enclave Bill" — that included provisions for release of the Census Bureau's Standard Statistical Establishment List (SSEL) to qualified Federal and state agencies for statistical purposes. This was only the latest in a long series of failures to achieve what was recognized by many nearly 50 years ago as a sensible goal: the use of shared business lists to reduce costs and increase the comparability of data from establishment surveys conducted by different agencies in our decentralized statistical system. It was our view in 1984 that it was time for the parties involved to engage in a constructive search for new solutions. In particular, we recommended consideration of alternatives such as the construction of a business list containing only information for units for which the Census Bureau or other cooperating agencies (**not** including IRS) had collected information directly.

Current Status

Following the withdrawal of administration support of the Enclave Proposal, the Census Bureau focussed its efforts on legislation aimed at a narrower goal: release of SSEL information to qualified Federal and state statistical agencies. However, because the SSEL incorporates some information from tax returns, the IRS strongly opposed amending the Tax Code to permit release of SSEL information to other agencies.

In March 1986, the Economic Policy Council established a Working Group on the Quality of Economic Statistics. The quality of business lists used in economic surveys was one of five issues to which the Working Group assigned its highest priority. In its April 1987 report to the Economic Council, the Working Group recommended a two-pronged approach to the improvement of business lists (Economic Policy Council, 1987):

(1) The Commerce Department should submit legislation to permit the Census Bureau to disclose business identification and classification information to specified statistical agencies. **No tax data received from IRS would be included in these disclosures.**

(2) Under the authority of the Paperwork Reduction Act, OMB should designate the Bureau of Labor Statistics (BLS) and the National Agriculture Statistics Service (NASS) as "Central Collection Agencies" for certain nonfarm and farm business lists, respectively.

The Working Group recommended that drafts of the legislative proposal and the administrative directives be "completed and coordinated" by June 1987. Both BLS and NASS have submitted draft proposals for carrying out their functions as designated central collection agencies, but the formal designations have not yet been made by OMB. As of November 1987, the Census Bureau had not yet submitted a legislative proposal.

The outcome of these new steps is difficult to predict. Much depends on: whether legislation permitting the Census Bureau to share the SSEL can be obtained; whether BLS and NASS can obtain the additional funds needed for their business lists; and whether BLS can work out satisfactory arrangements with the state employment security agencies that administer the Unemployment Insurance (UI) program and are the main source of BLS' business lists. In connection with the last of these points, it is encouraging to note that BLS has already obtained permission from about 40 States to share portions of the UI list file with NASS for the latter agency's use in list building, especially in the agricultural services sector. The real test will be whether arrangements can be developed for two-way list sharing between the Census Bureau and BLS and between the Census Bureau and NASS. Definitive information on how well this new system is working will come, at the earliest, with the 1992 Agriculture Census and Economic Censuses.

1.6 Standardization of reporting units for administrative and statistical data collections

Goal 6: Continue efforts to develop more consistent and compatible procedures for defining and identifying reporting units in both administrative and statistical data collections.

Status in 1984

This goal relates primarily, although not exclusively, to business or economic reporting units. We pointed out that businesses, especially those that are employers, are burdened by administrative reporting requirements (e.g., for information on employment and wages) that overlap, but are not fully compatible with respect to identifiers and definition of reporting units, and that are imposed by agencies that share only limited information with each other. Without better integration of these administrative reporting systems, it will be difficult to achieve standardization in the statistical programs that rely on them for the identification and classification of reporting units. (The paper by Colledge, 1987b, at this symposium, covers some of the same issues in a Canadian context.)

In 1984, a plan for an Establishment Reporting Unit Match Study (ERUMS) was being developed under the aegis of the Administrative Records Subcommittee, Federal Committee on Statistical Methodology (Cartwright et al., 1983; Buckler, 1985). The goal of ERUMS was to do a sample matching study of BLS and SSA records in one state, in order to identify differences in the coverage and content of the two agencies' systems, and to develop recommendations for achieving greater comparability.

Current Status

Progress in the ERUMS project has been slow, but a sample of matched and unmatched records from the two systems is now available and is undergoing detailed analysis. The

slow pace of this research effort has resulted, in part, from the legal and administrative intricacies that must be mastered in any project that links records from different agencies and, in part, from the fact that execution depends on agency staffs whose main priorities lie elsewhere. Nevertheless, the participants feel that the experience and the results will be useful, not only for what they show about the relation of the BLS and SSA business lists, but as a model for other similar studies. In particular, matching studies involving BLS, Census Bureau and NASS business lists could contribute valuable information to guide the proposed restructuring of business lists used for economic surveys by Federal statistical agencies (see subsection 1.5, above).

There has been no significant progress toward integration of the underlying administrative reporting systems. In one sense, the burden on employers has increased. In 1978, burden was reduced by switching from quarterly to annual reporting of individual employee earnings for social security purposes. More recent Federal legislation, however, requires that employers in all states report individual wages quarterly under the unemployment insurance system, whereas previously some states required reporting of individual wages only for employees who had submitted claims for unemployment benefits.

One step forward has been the inclusion of Employer Identification Numbers (EINs) in BLS' employer identification file. This addition has put BLS in a better position to fulfill its potential role as an OMB-designated supplier of business lists; for example, it will allow BLS to link records with lists from other sources.

1.7 Summary

Figure 4 summarizes our views about progress, or the lack of it, since 1984, in meeting the six goals set at that time. The subjectivity of this exercise must be recognized from the start. The goals were broadly defined: criteria for evaluating achievement are at most implied, not specifically defined. Other observers could easily take a more optimistic or pessimistic view than we have.

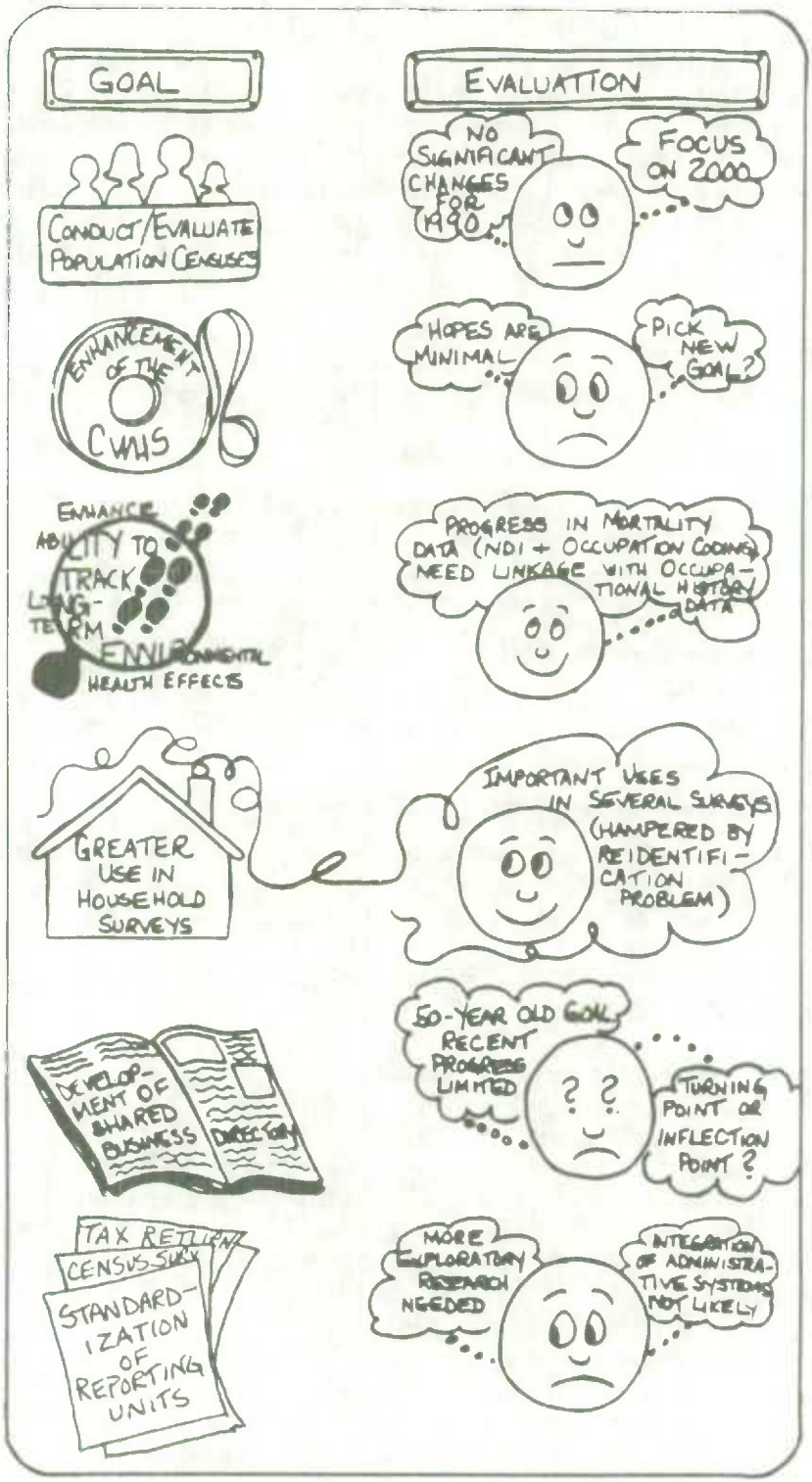
There should be a "?" after 2000 (Goal 1, Figure 4), which is the saddest comment we can make about progress towards an Administrative Record Census (ARC). What is done elsewhere in the next few years may make the difference, especially what gets done in Canada. To its credit the U.S. Census Bureau is not ignoring this possibility, but we feel that (and our colleagues from Census may want to disagree) the Bureau's primary focus has been elsewhere. The question of whether to adjust the 1990 decennial census counts for under (over) coverage has preoccupied the agency. The decision not to adjust in 1990 is a blow to an ARC in 2000 since, if the decision to adjust had been made, administrative records might have been the key resource for the local area figures.

The short-term prospects for achieving Goal 2, i.e., expanding the CWHS to become an accessible multi-purpose longitudinal data system, appear to be very low. Some limited successes seem possible; nonetheless, we have reluctantly suggested that attention be diverted to other administrative data systems which do not have all the attractive features of the CWHS, but offer better prospects for further development. At the other end of the spectrum is Goal 4, greater use of administrative records in household surveys. Here we see significant progress, with the one cloud on the horizon being the reidentification problem that inhibits the release of microdata files containing linked survey and administrative data.

For the other goals, we are probably no worse off than we were in 1984, but achievements since then are not any cause for celebration. The overall picture is not encouraging.

Figure 4

Progress in Reaching Goals: 1984-87



What are the factors inhibiting more effective statistical uses of administrative records, and is there anything we can do to change the ways in which they operate? These are the questions to be examined in the remaining two sections of this paper.

2. DETERMINANTS OF PROGRESS IN STATISTICAL USES OF ADMINISTRATIVE RECORDS

In this section we discuss a number of factors which we believe are especially critical to success in making cost-effective applications of administrative records to statistical programs. The most important success factor is an agreement on goals and this has not been achieved. We have no shared vision of the future, and at present, there is no effective process for obtaining such a vision or guiding its execution. In addition to this general problem, there are four specific factors that may also be worth commenting on and which may have parallels in Canada and elsewhere:

- **Changes in administrative record systems**

  How extensive are they and what mechanisms can be developed to adapt to them?

- **Laws and policies governing access and disclosure**

  These laws and policies affect initial access to administrative data by statistical agencies, the ability to link administrative data with census or survey data, and the extent to which aggregate data and microdata, derived wholly or in part from administrative sources, can be released for statistical uses.

- **Attitudes and perceptions about statistical uses of administrative data**

  It is essential to consider the attitudes and perceptions of all parties involved: the custodians of the administrative data systems, the statistical agencies that use the data, and the general public.

- **Coordination of relevant activities within the U.S. statistical system**

  The three previous elements are likely to be important in all countries that use administrative data for statistical purposes. This one, however, has special relevance in the United States because of its decentralized statistical system.

The remainder of this section looks at each factor in greater detail, focusing primarily on developments since mid-1984.

2.1 Changes in administrative record systems

The coverage, content, quality, and accessibility of administrative record systems are subject to frequent changes, depending on program requirements and resources available. When systems are expanded, significant new opportunities for statistical uses arise. When systems contract, deteriorate or become less accessible, ongoing statistical uses can be adversely affected. Concern about adverse effects was expressed by Butz (1984) in his comments on our six goals:

> *When administrative records are substituted for statistical data rather than complementing them, serious problems can and sometimes do arise. As a substitute, administrative records supplant statistical data. This can leave a statistical agency dangerously vulnerable to changes and failures outside its control and can expose users to potential loss of important data.*

Experience has shown that statistical users of administrative records must be willing to invest significant resources in monitoring system changes, if they want to avoid unpleasant surprises and take advantage of new opportunities. Active channels of communication between custodians and statistical users are a must, and users must be aggressive in establishing and maintaining such links. If they are successful in doing this, they may even be able to exert some influence on the kinds of changes that occur.

Since mid-1984, the most significant changes in administrative records systems of interest to statisticians are those resulting from passage of the Tax Reform Act of 1986. As mentioned in Section 1, resulting changes in the coverage and content of IRS and SSA administrative files have significant implications for both the IRS Statistics of Income (SOI) program and many of the uses of IRS/SSA records made by the Census Bureau (Jabine, 1987b). For example, the Statistics of Income Division of IRS is planning a major revision of the individual SOI sample. In two separate papers at this Conference, Gates (1987) and Hanczaryk and Jonas (1987) make clear that the Census Bureau has been active in evaluating the effect on its programs of changes resulting from tax reform. These papers illustrate some of the elements of the activist strategy needed by statistical agencies to keep abreast of changes in the content and quality of administrative record systems.

Another significant change has been the gradual development of Medicare program files to the point where they are more accessible for statistical uses, not just for coverage of the population 65 and over, but to enhance health-related survey data for persons in this age group. The Health Care Financing Administration has recently undertaken several analytical studies based on linkages of Medicare records with administrative records of other Federal and state agencies. Uses and costs of Medicare services on the last year of life, by cause of death, were studied by linking Medicare records with the National Death Index and the NCHS mortality file (Riley et al., 1987). Medicare utilization by disabled worker social security beneficiaries was studied by linking Medicare records with administrative files for SSA's disability program (Bye et al., 1987). The impact of nursing home care on total Medicare and Medicaid expenditures has been studied by linking Medicare and Medicaid administrative files for four states (McMillan et al., 1987).

Jabine (1984) pointed to the need for a more systematic effort to obtain information on prospective and actual changes in administrative record systems and disseminate that information to users throughout the U.S. statistical system. It is a pleasure to report that, building on the beginning made by Crane and Kleweno (1985), the Census Bureau has taken up this challenge and has established an Administrative Records Information System (ARIS), as reported by Gates in his paper for this Conference.

These developments are encouraging: we feel that the statistical agencies today are more alert to the problems and opportunities created by changes in administrative record systems and are doing a better job of monitoring changes.

2.2 Laws and policies governing access and disclosure

Statistical uses of administrative records in our decentralized statistical system are heavily dependent on laws, regulations and policies that govern disclosure of individually identifiable records. As noted in Section 1, disclosures of IRS and SSA administrative records, which have the greatest potential for statistical uses, are narrowly circumscribed by legislation and regulation. Uses of state-controlled administrative data, such as vital records and records maintained for the unemployment insurance system, require accommodation to a wide spectrum of state laws governing access to their records.

The same laws that govern transfer of identifiable records between agencies also require, in nearly all cases, that there be no disclosure of identifiable information associated with releases of aggregate data or public-use microdata files. This

requirement is always stated in absolute terms — no disclosure — although most statisticians will admit that no data of any consequence can be released without incurring some risk of disclosure (e.g., Paass, 1985; Cox et al., 1985; Duncan and Lambert, 1987). Such risks are not easy to quantify, but there is a general feeling that they are increasing because of the greater availability of data about identifiable individuals and the development of more sophisticated record-linkage techniques.

Another important question is: what should those who supply information be told about planned or potential uses of that information for statistical purposes? Should those who file tax returns be told about all planned statistical uses of the information they provide? What should survey respondents be told about planned linkages of their survey data to information about them that already exists in administrative record systems?

These issues were already prominent by 1984 (American Statistical Association, 1977) and have been receiving increased attention since then (e.g., Gastwirth, 1986; Jabine, 1986; Scheuren, 1986). Several developments are worth noting:

- As mentioned in subsection 1.5, efforts to develop legislation to support business list-sharing arrangements are now proceeding under the assumption that no data received directly from IRS by the Census Bureau will be redisclosed to other agencies.
- In 1985, the National Agriculture Statistics Service (NASS) obtained legislation that provides strong protection for the confidentiality of data NASS collects in surveys, but does not preclude disclosure of such information to other agencies for statistical purposes.
- In 1986, after extended negotiations, the National Cancer Institute was able to develop arrangements to obtain SSNs from the Social Security Administration to transmit to IRS (through the National Institute for Occupational Safety and Health), in order to obtain current addresses for persons being traced in epidemiological followup studies. The key to this arrangement was agreement by IRS that it would use any data provided by NIOSH solely for the purpose of accessing current address information, i.e., not for any compliance activities.
- As mentioned in subsection 1.4, the Bureau of the Census has adopted a more restrictive policy concerning release of public-use microdata files that contain administrative record data linked to Census or survey data collected under the authority of Title 13.
- There appears to be a growing consensus, among both data producers and data users, that there is a need for mechanisms other than public-use files to allow researchers access to microdata files that do not qualify for unrestricted access. Two approaches are being mentioned: restricted release arrangements that provide severe penalties for any users who disclose identifiable data (and compensation for persons harmed by disclosure of information about them) and interactive access to data files in the custody of the data producers, with screening of outputs to avoid disclosure.
- The Committee on National Statistics and the Social Science Research Council have been jointly active in seeking solutions to confidentiality and data access problems. A Conference on Access to Public Data was convened in November 1985 (Pearson, 1986) and, in September 1987, a workshop was organized to discuss access and confidentiality issues associated with a proposed followup study based on SSA's Longitudinal Retirement History Survey. A comprehensive panel study on Confidentiality and Data Access will be undertaken starting in mid-1988. Although the scope of the panel study is not limited to statistics based on administrative records, the issues to be studied will be highly relevant to statistical uses of administrative records.

Some of the events that have already occurred are positive in terms of facilitating statistical use of administrative records, but they represent minor progress at best. Looking at each of the six goals discussed in Section 1 of this paper, one has to conclude that except for Goal 1 (use of administrative records for decennial censuses and current population estimates), there are legal restrictions to achieving some of the desired results in every case. Laws can, of course, be changed, but prospects for new legislation are strongly affected by attitudes and perceptions of data custodians, data users and the general public. These attitudes and perceptions are the subject of the next subsection.

2.3 Attitudes and perceptions about statistical uses of administrative records

In the most direct sense, whether we move toward increased statistical use of administrative records depends on the institutional attitudes of agencies like IRS and SSA that are custodians of major administrative record systems and agencies like the Census Bureau that are the primary statistical users of administrative records.

Since passage of the Tax Reform Act of 1976 (and even earlier), the position of IRS has been to resist all expansion of non-tax uses of tax return data, whether for statistical or non-statistical purposes. Their justification for this position is that any such expansion may adversely affect tax compliance. This concern led directly to the alternative method of developing business lists, now being pursued, that would exclude the direct use of IRS data (see Section 1.5). Indeed, it is the personal and professional position of the second author of this paper that broader statutory access to tax records is undesirable. Other alternatives need to be pursued first.

The Census Bureau's position toward statistical uses of IRS data is ambivalent. As described in Gates' paper for this Symposium, Census makes substantial use of IRS administrative records for economic censuses and surveys, current population estimates and evaluation of coverage in the Census of Population. Such uses have been fully described in technical meetings and reports. On the other hand, until recently the Census Bureau has resisted making any statements to survey respondents about possible linkages of their survey data to information about them obtained from IRS. Even though identifiable records flow in only one direction, from IRS to the Census Bureau, there is a concern that linkage of the two agencies in the public's mind might diminish the Census Bureau's ability to achieve high response rates in censuses and surveys. Census staff point to the strong opposition that developed to censuses in Germany and the Netherlands, and they attach great importance to avoiding any activities that might provoke similar reactions in this country (Butz 1984, 1985b).

Thus, the attitudes of the custodians of administrative records and the agencies that use them for statistical purposes depend, at least in part, on their views about how the public might react to highly visible, widely publicized statistical uses of administratives records. Knowledge of public attitudes and perceptions about statistical uses of administrative records can be obtained in two ways: by reviewing manifestations of such attitudes as reported in the media, and by making direct inquiries in public opinion surveys.

Since 1984, perhaps the most directly relevant episode reported by the media has been the public furor in Sweden leading to the termination of data collection for a longitudinal research data base that was being maintained for a social science research project named Metropolit. The data base linked longitudinal data from surveys and several administrative record sources for all persons born in the Stockholm metropolitan area in 1953 and living there when the data base was established 10 years later. The full story is too involved to tell here, but when many of the persons included in the data base became fully aware of its content and scope, their reaction led to the removal of identifiers from the data base and, hence, the preclusion of further linkages. Of particular importance in the context of this paper is the fact that when the controversy broke, cooperation rates in

the Swedish Labor Force survey declined by about 5 percentage points and have not since returned to their previous level (Dalenius, 1986).

The use of surveys to learn directly about public attitudes toward statistical uses of administrative records is difficult because the issue is not of much interest to most members of the public, and few persons have more than the vaguest idea of what is presently being done. Questions on the subject of non-tax uses of IRS data that were included in surveys sponsored by IRS in 1984 and 1986 led to somewhat inconsistent results, perhaps because the lack of salience of this topic for most respondents made their responses sensitive to differences between the two surveys in the way the questions were framed and administered (Gonzalez and Scheuren, 1985; Scheuren, 1986).

For the 1987 IRS Taxpayer Attitudes Survey, the approach to investigation of this topic was modified. The questions on this topic (97 through 99) were arranged in the following sequence (see the appendix to the remarks by Tom Jabine, 1987a, as a panelist at the concluding session of this conference):

- An initial statement introduces the topic of data sharing among different government agencies. A concrete illustration is included.
- Q. 97, which uses a five-point anchored scale, asks for views about 6 specific considerations that might cause people to favor or oppose data sharing. Actually, there are only 5 considerations: as a test of internal consistency Q. 97f states the same considerations as Q. 97c, but in the opposite form.
- Q. 98, coming after respondents have been exposed to various relevant considerations, now asks respondents for their overall view on data sharing. Q. 98b allows for open responses justifying the response selected in Q. 98a.
- Q. 99 asks respondents for their views on 4 specific non-tax uses of IRS data. Two of the uses (a and d) are statistical and 2 are non-statistical. In each case, both the receiving agency and the purpose for sharing are identified. The order in which the items were presented to respondents varied.

The weighted overall percent distributions of responses to these questions are also shown in the same appendix (Jabine, 1987a). The target population consisted of persons who normally file Federal tax returns, with the further specification that joint filers were to be represented by the taxpayer considered to be the more knowledgeable about the return. The most interesting findings are:

- A substantial majority want to know which agencies have information about them and why they want it.
- Smaller majorities agree that data sharing would reduce public response burden and the cost to the government of getting the information it needs.
- Overall, the proportion of respondents opposing data sharing is slightly larger than the proportion who favor it (41 percent vs. 38 percent). Strong opposition exceeds strong approval by a considerable amount.
- Reactions to data sharing for specific purposes vary within a relatively narrow range. Release to the Justice Department for criminal investigations receives the highest approval and release to state governments for improving their tax collections is opposed by the largest proportion of respondents. Statistical uses by the Census Bureau and the Commerce Department ranked between the two non-statistical uses.
- When faced with specific instances of data sharing rather than the general concept, fewer respondents are undecided and reactions are somewhat more favorable.

Data of this kind are helpful in understanding public attitudes toward data sharing, and we are in the process of undertaking a more detailed analysis. Microdata files from the survey will be available to researchers through the Inter-university Consortium for Political and Social Research. However, we must caution that public attitudes on these subjects are notoriously fickle. In the campaign for this year's general election in Australia, the government party told voters that it wanted to introduce a national identity card. Opinion surveys showed that more than 60 percent of Australians liked the idea. However, subsequent to the election, articulate opposition arose and surveys showed that support for the idea was nearly cut in half (**Washington Post**, 1987).

2.4 Coordination of relevant activities within the United States statistical system

The present lack of strong central coordination of our decentralized statistical system makes it difficult for statistical agencies to exert much influence on the coverage and content of administrative record systems, or on the conditions of access to administrative data sources for statistical applications. The Office of Statistical Policy (OSP), in the Office of Management and Budget, is formally charged with coordination of Federal statistical activities. With limited resources, OSP focuses most of its attention on control of response burden, maintenance of standard classifications, such as the Metropolitan Statistical Areas, and interagency collaboration on technical issues through the Federal Committee on Statistical Methodology (FCSM). As described in our 1985 paper, the Administrative Records Subcommittee of the FCSM has sponsored several activities aimed at exchange of information on technical aspects of statistical uses of administrative records. These activities have continued, although at a reduced level, since 1984.

The Office of Statistical Policy does not, however, attempt to lead the Federal statistical agencies in any efforts to engage in intermediate or long-term planning for the entire statistical system. As we said in our 1985 paper:

> *Many individual statistical agencies are doing an excellent job of long-range strategic planning. However, they do it in the context of their own functions, programs, and interests, and not as part of an overall statistical system designed to meet the information needs of the government and the public in the most efficient way possible. As a result, we believe there is a bias that favors direct data collection programs under the control of a single agency in preference to those making greater use of administrative records, but requiring sharing of data and close cooperation by two or more agencies. This situation explains the relative scarcity of resources for research and development work on new uses of administrative records, and it may also account for the failure of efforts to obtain legislation needed to implement a shared business directory and other projects that require exchanges of identifiable data.*

The situation is not much different today.

When such a vacuum exists, various groups do what they can to fill it. We list here some activities of both governmental and non-governmental organizations since 1984 that have had or may be expected to have some influence on statistical uses of administrative records.

Governmental

- At the policy level, we have already mentioned the Economic Policy Council's Working Group on the Quality of Economic Statistics, and its recommendations on sharing of

business lists (subsection 1.5). It may be noted that the agencies involved are already behind the recommended schedule for implementation. Other recommendations of the working group, especially those relative to improvement of merchandise trade statistics, may also affect significant statistical uses of administrative records, although it is not always clear whether they will lead to greater or less dependence on administrative sources.

- We have also expressed our support for the recent establishment of an Administrative Record Information System by the Census Bureau. Another new Census Bureau activity, which has facilitated exchange of information at the technical level, is its series of annual research conferences, started in 1985. The Proceedings of these conferences have been invaluable (e.g., Bureau of the Census, 1987). Several of the presentations at these conferences have been relevant to uses of administrative records. In the 1987 conference, for example, Tippett (1987) and Diemer (1987) discussed the use of administrative sources to evaluate the quality of data from housing censuses and surveys; and Colledge (1987a) described the role of tax data in a major redesign of Statistics Canada's business surveys.
- The IRS has continued regular publication of collected papers, presented at the annual meetings of the American Statistical Association, on the subject of statistical uses of administrative records (e.g., Internal Revenue Service, 1987b).
- Since 1984, the Committee on National Statistics has continued, through panel studies and other activities sponsored primarily by Federal agencies, to address policy and technical issues of broad interest to the U.S. statistical system. Many of these activities involve statistical uses of administrative records. In subsection 1.1, we mentioned the Panel on Decennial Census Methodology. One of the three major issues this panel was asked to investigate was "Uses of administrative records, including investigation of the possible utility of various types of records for improving the accuracy of census counts, and the efficiency of census operations" (National Research Council, 1985). Two new panel studies are just getting underway, one on foreign trade statistics and one on confidentiality and data access. In both of these studies, issues related to statistical uses of administrative records are likely to be an important component.

Nongovernmental

- The Council of Professional Associations on Federal Statistics (COPAFS), has provided a forum for agency policymakers to meet with each other and with representatives of professional societies to discuss statistical policy matters of current interest. COPAFS has effective links with the U.S. Congress and serves as a clearinghouse for information on congressional actions affecting the statistical agencies.
- In 1985, the National Association of Business Economists formed a Statistics Committee "to act as a liaison between the membership of NABE and the Federal, state and local statistical communities, to help insure the integrity, scope, accuracy and timeliness of business statistics" (National Association of Business Economists, 1987). One of the committee's concerns has been the quality of business lists used in economic surveys.
- The American Economic Association has formed a committee, chaired by Thomas Juster, to review the adequacy of the concepts, definitions and classification systems used in our economic statistics, in the context of recent structural changes in our economic and social institutions.

Activities like these are important to the health of the U.S. Federal statistical community and deserve support and encouragement. They are, however, far from enough,

in and of themselves, to move the U.S. Federal statistical agencies towards a coherent overall strategy on the statistical uses of administrative records (or indeed on many other cross-cutting issues).

3. ANOTHER FORECAST

Obviously, we can only guess at the answer to the second question in the title of this paper: Where are we going? One thing seems fairly certain: the technological developments of the information age will continue at a rapid rate. These developments will bring substantial opportunities for wider and more effective statistical uses of administrative records.

How can the U.S. statistical system avail itself of these opportunities? As an aid to thinking about this question, it may help to look at the statistical use of administrative records as a **system.** Every system, whether statistical or not, has three basic elements: inputs, the processes applied to the inputs, and outputs. Figure 5 displays what we consider to be a few of the main elements in this particular system.

Inputs

In our earlier analyses, we have focused our attention, perhaps unduly, on administrative records as inputs. They are, of course, the **sine qua non;** and the monitoring of changes in their coverage, content, quality and conditions of access — as is now being systematically carried out by the Census Bureau in its Administrative Records Information System — is critical to the entire enterprise.

Nevertheless, other inputs are of equal or even greater importance. Statistical agencies must have computers capable of handling very large, complex data bases. Increasing the number of skilled people with a real commitment to excellence is also essential. These people will require state-of-the-art familiarity with the tools shown in the "processes" column of Figure 5: statistics, quality control, telecommunications, and eventually artificial intelligence (AI) approaches.

Processes

A key feature of the information processing that converts administrative record inputs to statistical outputs is **record linkage.** Rapid strides are being made in the development of model-based multipurpose record-linkage systems (Internal Revenue Service, 1985; Jabine and Scheuren, 1986). The underlying theory, thanks to the early work of Fellegi and others, seems well understood (Kirkendall, 1985); nonetheless, its application requires skills and experience that are in short supply. Further research, knowledge-sharing and development of user-friendly record-linkage systems are all needed (see Internal Revenue Service, 1985, Recommendations, pp. 3-4). This Conference is itself greatly advancing that goal as exhibited by the many papers being given in this area.

The organizational framework within which these processes operate is not likely to change much. Our decentralized statistical system might be likened to a distributed computer-processing environment with a lot of mainframes (not just one huge system) linked together with mini-computers and PCs to form a network. The trouble is, to continue the analogy, that we don't have an adequate level of **connectivity.** We badly need a system-wide strategic planning process to drive resource allocation decisions (especially for research and development) for Federal statistical programs. Turf boundary issues are distracting us from our main mission as public servants. Differing points of view are healthy and necessary for real progress, but we need better ways of resolving our differences. Some problems, for example the development of shared business lists (as noted above), are 50 years old and still not resolved. Such weaknesses should not continue to be tolerated.

Figure 5

Looking at Statistical Uses of Administrative Records as a System: Some Examples

Inputs	Processes	Outputs
Administrative records Computers and peripherals Agency personnel	**Direct:** Systems design Information processing **Framework:** Organizational structure Strategic plan **Tools:** Statistical science Quality Control and improvement Expert systems, AI Telecommunications	**Products:** Publications **Data files:** Aggregate data Microdata **Customers:** Governments Businesses Universities Individual citizens

Outputs

More attention is needed to the output side of the system. We need to ask ourselves whether we are sufficiently "Customer Driven". In our view, more outreach is needed in U.S. statistical programs, especially in the area of administrative records. We are supposedly in the "Information Age," yet many of the products we produce and their timing have not changed enough to adequately reflect this. In particular, because of confidentiality constraints, we need to look for alternatives to unrestricted public-use data files that will allow users sufficient access to needed data without compromising the confidentiality of the individual information.

Final Comments

There are some obvious dimensions for future progress. U.S. statistical programs should benefit, for example, from the major improvements being attempted by a number of administrative agencies, especially the Internal Revenue Service. At IRS, we are very, very serious about doing a better job for the American people (Scheuren, 1987). This should lead not only to improved outputs but also to improved technologies. The use, by administrative agencies, of artificial intelligence and other computer-intensive approaches, particularly in the area of expert systems (Beckman, 1987), should help in building the methodological infrastructure needed for still greater statistical uses of administrative records. The telecommunications revolution will also be a factor, perhaps leading eventually at IRS to such things as widespread use of electronic filing of tax returns. (In 1988, there may be as many as half million returns filed electronically.) Eventually, a completely "return-free" system may exist, with all the improvements in quality and timeliness that can go along with such a system (Wedick, 1986; Internal Revenue Service, 1987a).

In other words, without any real effort by them, statistical agencies will benefit from a number of forces operating in administrative departments. To take full advantage of these forces, however, there is some need for improving the organizational structures that exist to do government statistics (especially in the direction of better planning and coordination of R&D resource allocation decisions).

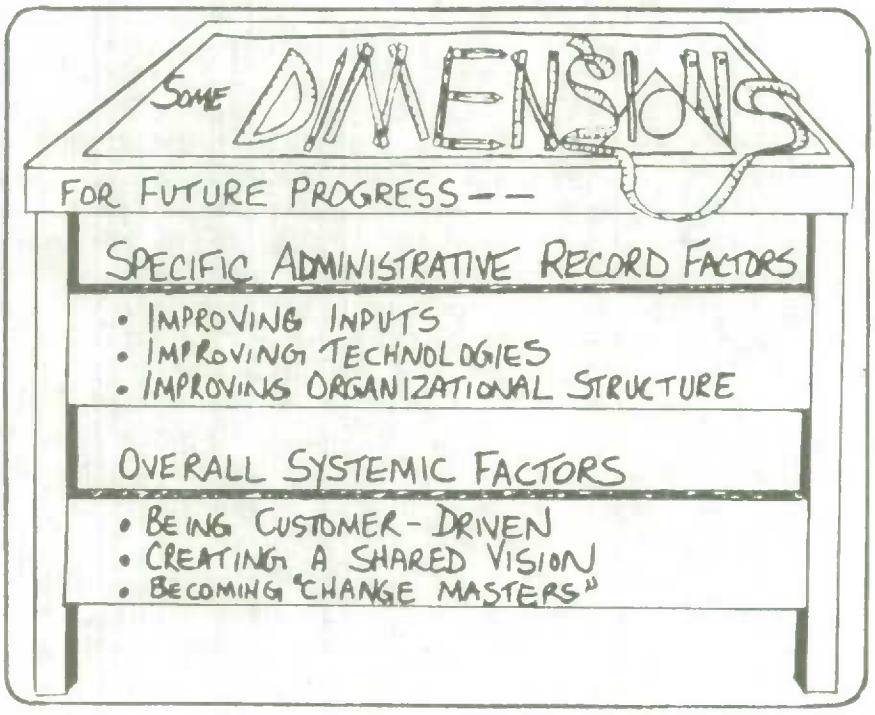
Finally, we would like to point to three overall systemic requirements for progress (see Figure 6). In a sense we have already discussed the first two of these. If the U.S. is to make better statistical use of its administrative records, all of us, both custodians like the SOI Division at IRS and statistical agency employees like our friends at this Conference from the Census Bureau, need to let ourselves be held accountable by our customers. We don't just mean our ultimate customers, the American people, but all the intermediate customers, too. In fact, we are each others' customers, as Figure 1 implies.

As Peters has said, we need to learn to listen "naively" to our customers (Peters and Waterman, 1982). Naturally, too, there are many different customers and being "customer-driven" can literally mean being driven crazy if there is no mechanism to resolve conflicts. Perhaps the best mechanism for conflict resolution is the creation of a shared vision. A strategic planning process is one of the tools that has already been mentioned and which may be useful here. Planning together is not enough, though. Developing a shared vision requires a lot of hard work in sorting out values; the impetus for this effort doesn't exist right now. Each of us seems to have more than enough to worry about without trying to tackle problems that have seemingly lasted forever. We're not ready to despair here, but there is no point in listing any pat answers.

Our final observation is that we have two options in the way we deal with the future: to become "Change Masters" (Kanter, 1983) or to be mastered by change. More use of administrative records for statistical purposes is inevitable. Whether we welcome this

development or not, it will come like a glacier or an avalanche. To return to a point made at the beginning: Not only are we the people **who do what is obvious but not easy,** but we are also the people **who work in a field which inevitably is growing in scope and challenge.** There is an old Chinese curse which roughly translated goes, "May you live in interesting times." Well, there is no doubt that in this field we live in interesting times. Whether or not that is a curse is at least partly up to us.

Figure 6



ACKNOWLEDGMENTS

The authors would like to thank Bettye Jamerson, Wendy Alvey and especially Beth Kilss for their assistance in the preparation of this paper. Thanks are also due to Nancy Dutton, Dorothy Farmer and Darlene Reynolds for typing assistance on the many drafts of the paper and speech.

REFERENCES

Alvey, W., and Scheuren, F. (1982). "Background for an Administrative Record Census," *Proceedings of the American Statistical Association, Social Statistics Section,* 137-146.

American Statistical Association (1977). "Report of Ad Hoc Committee on Privacy and Confidentiality," *The American Statistician,* vol. 31, 59-78.

Beckman, T. (1987). "Trends in Selection and Development of Applications Using Artificial Intelligence Technology," unpublished IRS working paper.

Bowie, C., and Kasprzyk, D. (1987). "The Use of Administrative Records with Data from the Survey of Income and Program Participation," presented at the International Symposium on Statistical Uses of Administrative Data, Statistics Canada, Ottawa, Ontario.

Brackstone, G.J. (1987a). "Issues in the Use of Administrative Records for Statistical Purposes," *Survey Methodology*, vol. 13, no. 1, 29-43.

Brackstone, G.J. (1987b). "Statistical Uses of Administrative Data: Issues and Challenges," paper presented at the International Symposium on Statistical Uses of Administrative Data, Statistics Canada, Ottawa, Ontario.

Bristol, R.B., Jr. (1985). "Tax Modelling and the Policy Environment of the 1990's," *Multi-National Tax Modelling Symposium Proceedings*, Ottawa, Ontario, II-11-II-17.

Buckler, W. (1985). "Employer Reporting Unit Match Study (ERUMS): A Progress Report," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 434-437.

Buckler, W., and Smith, C. (1980). "The Continuous Work History Sample (CWHS): Description and Contents," *Economic and Demographic Statistics*, Social Security Administration, 165-174.

Bureau of the Census (1983). "Feasibility of an Administrative Record Census in 1990," unpublished report of the Subcommittee on an Administrative Records Census, Committee on the Use of Administrative Records in the 1990 Census, U.S. Department of Commerce.

Bureau of the Census (1987). *Proceedings of the Third Annual Research Conference*, U.S. Department of Commerce.

Butz, W. (1984). "The Future of Administrative Records in the Census Bureau's Demographic Activities," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 61-63.

Butz, W. (1985a). "Comment: The Future of Administrative Records in the Census Bureau's Demographic Activities," *Journal of Business and Economic Statistics*, vol. 3, no. 4, 393-395.

Butz, W. (1985b). "Data Confidentiality and Public Perceptions: The Case of the European Censuses," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 90-97.

Bye, B., Riley, G., and Lubitz, J. (1987). "Medicare Utilization by Disabled-Worker Beneficiaries: A Longitudinal Analysis," *Social Security Bulletin*, Vol. 50, no. 12, December.

Carroll, J. (1985). "Comment: Uses of Administrative Records: A Social Security Point of View," *Journal of Business and Economic Statistics*, vol. 3, no. 4, 396-397.

Cartwright, D., Levine, B., and Buckler, W. (1983). "An Update on Establishment Reporting Issues: Practical Considerations," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 481-486.

Childers, D., and Hogan, H. (1984). "Matching IRS Records to Census Records: Some Problems and Results," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 301-306.

Colledge, M. (1987a). "The Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada," *Proceedings of the Census Bureau Third Annual Research Conference*, Bureau of the Census, 550-576.

Colledge, M. (1987b). "Uses of Administrative Data in the Business Survey Redesign Project," presented at the International Symposium on Statistical Uses of Administrative Data, Statistics Canada, Ottawa, Ontario.

Cox, L., and Boruch, R. (1985). "Emerging Policy Issues in Record Linkage and Privacy," A paper presented at the 45th Meeting of the International Statistical Institute in Amsterdam, The Netherlands, August 12-22, 1985.

Cox, L., Johnson, B., McDonald, S., Nelson, D., and Vazquez, V. (1985). "Confidentiality Issues at the Census Bureau," *Proceedings of the Census Bureau First Annual Research Conference*, 199-218.

Crabbe, P., Sailer, P., and Kilss, B. (1983). "Occupation Data from Tax Returns: A Progress Report," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 312-317.

Crane, J., and Kleweno, D. (1985). "Project LINK-LINK: An Interactive Database of Administrative Record Linkage Studies," *Record Linkage Techniques 1985*, Internal Revenue Service, 311-315.

Crouse, W., Schuster, L., Rosenberg, H., Kametani, D., and Sestito, J. (1983). "Using the Census Bureau's Occupation and Industry Coding System for Coding Death Certificates," Proceedings of the American Statistical Association, Section on Survey Research Methods, 306-311.

Czajka, J. (1987). "Turning the Tables: Imputing for Item Nonresponse When Donors are Scarce," paper presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Ontario.

Dalenius, T. (1986). "The 1986 Invasion of Privacy Debate in Sweden," unpublished report.

Diemer, W. (1987). "Micro-Evaluation of the 1980 Census of Housing," *Proceedings of the Census Bureau Third Annual Research Conference*, Bureau of the Census, 437-476.

Duleep, H. (1986). "Incorporating Longitudinal Aspects into Mortality Research Using Social Security Administrative Record Data," *Journal of Economic and Social Measurement*, vol. 14, 121-133.

Duncan, G., and Lambert, D. (1987). "The Risk of Disclosure for Microdata," *Proceedings of the Census Bureau Third Annual Research Conference*, 263-274.

Economic Policy Council (1987). "Report of the Working Group on the Quality of Economic Statistics," U.S. Department of Commerce, Washington, D.C.

Energy Information Administration (1987). *Residential Energy Consumption Survey: Consumption and Expenditures, April 1984 Through March 1985*, Part I: National Data, Appendix A, How the Survey was Conducted, U.S. Department of Energy.

Fay, R.E., and Herriot, R. (1979). "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association*, vol. 74, no. 366, 269-277.

Fellegi, I.P., and Sunter, A. B. (1969). "A Theory for Record Linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, 1183-1210.

Flaherty, D. (1979). *Privacy and Government Data Banks: An International Perspective*, Mansell Publications, London, U.K.

Gastwirth, J. (1986). "Ethical Issues in Access to and Linkage of Data Collected by Government Agencies," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 6-13.

Gates, G. (1987). "Evaluating the Effects of Tax Reform on Census Bureau Programs," paper presented at the International Symposium on Statistical Uses of Administrative Data, Statistics Canada, Ottawa, Ontario.

Gonzalez, M., and Hoza, C. (1978). "Small-Area Estimation with Application to Unemployment and Housing Estimates," *Journal of the American Statistical Association*, vol. 73, no. 361, 7-15.

Gonzalez, M., and Scheuren, F. (1985). "Future Work by the Conference of European Statisticians on Population and Housing Censuses," presented before the Thirty-third Plenary Session of the U.N. Conference of European Statisticians.

Hanczaryk, P., and Jonas, J. (1987). "Automated Quality Assurance Processing of Administrative Records Files," paper presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Ontario.

Hidiroglou, M.A., Morry, M., Dagum, E.B., Rao, J.N.K., and Särndal, C.E. (1984). "Evaluation of Alternative Small Area Estimators Using Administrative Data," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 307-313.

Hinkins, S., Jones, H., and Scheuren, F. (1987). "Updating Tax Return Selection Probabilities in the Corporate Statistics of Income Program," paper presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Ontario.

Hostetter, F., McCann, C., and Zirger, B. (1987). "Corporate Income Tax Records Used for Tax Policy Analysis," paper presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Ontario.

Hostetter, S. (1987). "Measuring Income for Developing and Reviewing Individual Tax Law Changes: Exploration of Alternative Concepts," *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Howe, G.R., and Spasoff, R.A. (Eds.). (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, University of Toronto Press.

Internal Revenue Service (1984). *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, vol. 2, Section VIII, "Summary and Prospects for the Future," U.S. Department of the Treasury, 651-680.

Internal Revenue Service (1985). *Record Linkage Techniques —1985*, Proceedings of the Workshop on Exact Matching Methodologies, U.S. Department of the Treasury.

Internal Revenue Service (1987a). "Report to Congress on the Return-Free Tax System," U.S. Department of the Treasury.

Internal Revenue Service (1987b). *Statistics of Income and Related Administrative Record Research: 1986-1987*, U.S. Department of the Treasury.

Jabine, T. (1984). "Proposal for an Administrative Records Monitoring System," in *Statistical Uses of Administrative Records: Recent Research and Present Prospects*, (vol. 1), Internal Revenue Service, U.S. Department of the Treasury, 37-38.

Jabine, T. (1986). "Selected Guidelines for Notification to Survey Participants", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1-5.

Jabine, T. (1987a). "Remarks at Panel Discussion," presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Ontario.

Jabine, T. (1987b). "Statistical Uses of Administrative Records in the United States: Some Recent Developments," presented at the Annual Meeting of the Statistical Society of Canada, Quebec.

Jabine, T., and Scheuren, F. (1985). "Goals for Statistical Uses of Administrative Records: The Next Ten Years," *Journal of Business and Economic Statistics*, vol. 3, no. 4, 380-391.

Jabine, T., and Scheuren, F. (1986). "Record Linkages for Statistical Purposes: Methodological Issues," *Journal of Official Statistics*, vol. 2, no. 3, 255-277.

Johnson, N., Rogot, E., Glover, C., Sorlie, P., and McMillen, M. (1985). "General Mortality Among Selected Census Bureau Sample Cohorts 1979-1981," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 428-433.

Kanter, R. (1983). *The Change Masters*, Simon and Schuster.

Kestenbaum, B. (1986). "An Accounting of the 1919 Birth Cohort," *Proceedings of the American Statistical Association, Social Statistics Section*, 397-400.

Kestenbaum, B. (1985). "The Measurement of Early Retirement," *Journal of the American Statistical Association*, vol. 80, no. 389, 38-45.

Kestenbaum, B., and Diez, G. (1981). "Geographic Mobility of Older Workers," *Proceedings of the American Statistical Association, Social Statistics Section.*

Kilss, B., Scheuren, F., and Buckler, W. (1980). "Goals and Plans for a Linked Administrative Statistical Sample," Proceedings of the American Statistical Association, Section on Survey Research Methods, 450-455.

Kirkendall, N. (1985). "Weights in Computer Matching: Applications and An Information Theoretic Point of View," *Record Linkage Techniques -- 1985*, Internal Revenue Service, 189-197.

McMillan, A., Gornick, M., Howell, E., Prihoda, R., Rabey, L., Russell, D., and Lubitz, J. (1987). "The Dually Entitled Medicare and Medicaid Elderly: Impact of Nursing Home Care on Costs," draft paper, Health Care Financing Administration.

Moore, J., and Marquis, K. (1987). "Using Administrative Record Data to Evaluate the Quality of Survey Estimates," presented at the International Symposium on Statistical Uses of Administrative Data, Statistics Canada, Ottawa, Ontario.

National Association of Business Economists (1987). "Report of the Statistics Committee."

National Research Council (1984). "Planning the 1990 Census: Priorities for Research and Testing," Interim Report of the Panel on Decennial Census Methodology, Committee on National Statistics.

National Research Council (1985). *The Bicentennial Census: New Directions for Methodology in 1990*, Report of the Panel on Decennial Census Methodology, Committee on National Statistics, Washington, D.C.: National Academy Press.

Newcombe, H.B. (1967). "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," *American Journal of Human Genetics*, University of Chicago Press, vol. 19, no. 3, Part I (May).

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959). "Automatic Linkage of Vital Records," *Science*, vol. 130, no. 3381, 954-959.

New York Times (1987). "U.S. Rejects Pleas to Adjust 1990 Census for Undercount," October 31.

Norwood, J. (1985). "Comment: Administrative Statistics: A BLS Perspective," *Journal of Business and Economic Statistics*, vol.3, no.4, 398-400.

Paass, G. (1985). "Disclosure Risk and Disclosure Avoidance for Microdata," paper presented at the International Association for Social Service Information and Technology.

Passel, J., and Word, D. (1980). "Constructing the List of Spanish Surnames for the 1980 Census: An Application of Bayes Theorem," paper presented at the annual meeting of the Population Association of America, Denver, Colorado.

Passel, J., and Word, D. (1987). "Problems in Analyzing Race and Hispanic Origin Data from the 1980 Census: Solutions Based on Constructing Consistent Populations from Micro-level Data," paper presented at the annual meeting of the Population Association of American, Chicago, Illinois.

Patterson, J., and Bilgrad, R. (1985). "The National Death Index Experience: 1981-1985," in *Record Linkage Techniques—1985*, Internal Revenue Service, U.S. Department of the Treasury, 245-254.

Pearson, R. (1986). "Researchers' Access to U.S. Federal Statistics," *Items*, vol. 41, nos. 1/2, 6-11.

Peters, T.J., and Waterman, R.H., Jr. (1982). *In Search of Excellence*, Warner Books, New York, New York.

Revenue Canada, Taxation (1986). *Multi-National Tax Modelling Symposium Proceedings* of the Canadian-U.S. Symposium on Tax Modelling held September 17-19, 1985, Mont Ste. Marie, Lac Ste. Marie, Quebec.

Riley, G., Lubitz, J., Prihoda, R., and Robey, E. (1987). "The Use and Cost of Medicare Services by Cause of Death," *Inquiry*, vol. 24, 233-244.

Rogot, E., Schwartz, S., O'Conor, K., and Olsen, C. (1983). "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 319-324.

Rogot, E., Sorlie, P., Johnson, N., Glover, C., and Makuc, D. (1985). "Mortality by Cause of Death Among Selected Census Bureau Sample Cohorts for 1979-1981," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 444-449.

Scheuren, F. (1985). "Methodologic Issues in Linkage of Multiple Data Bases," *Record Linkage Techniques -- 1985*, Internal Revenue Service, 155-178.

Scheuren, F. (1986). "Record Linkages for Statistical Purposes in the United States," *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, Ottawa, Ontario, 198-210.

Scheuren, F. (1987). "Notes on IRS Quality Improvement Process," speech presented at the OMB meeting of Departmental Productivity Officers, November 4, 1987.

Strudler, M., Oh, H.L., and Scheuren, F. (1986). "Protection of Taxpayer Confidentiality with Respect to the Tax Model," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 375-381.

Tippett, J. (1987). "Housing Data: The Quality of Selected Items," *Proceedings of the Census Bureau Third Annual Research Conference*, Bureau of the Census, 417-436.

U.S. Department of Health, Education and Welfare (1978). *Policy Analysis with Social Security Research Files*, Proceedings of a Workshop held March 1978 at Williamsburg, Virginia, Social Security Administration, Research Report No. 52, HEW Publication No. (SSA) 79-11808.

Waite, C.A. (1985). "Comment: The Future of Administrative Records in the Economic Programs of the Census Bureau," *Journal of Business and Economic Statistics*, vol. 3, no. 4, 400-401.

Wall Street Journal (1987). "Backers of an Adjusted Census Won't Take No for an Answer," November 3.

Washington Post (1987). "Australians in Uproar Over ID Cards," October 29.

Wedick, J.L., Jr. (1986). "Electronic Filing at the IRS: The Goal is Global," *Journal of Accountancy*, 110-116.

Wilk, M. (1985). "The Relationship Between Statisticians and Statisticians," *Survey Methodology*, vol. 11, no. 2, 89-94.

Wilson, R. (1983). "Postal ZIP Code Area Statistics from Internal Revenue Records," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 367-371.

Wolfson, M., Gribble, S., Bordt, M., Murphy, B., and Rowe, G. (1987). "The Social Policy Simulation Data Base: An Example of Survey and Administrative Data Integration," paper presented at the International Symposium on Statistical Uses of Administrative Data, Ottawa, Ontario.

SESSION II: INVITED PAPERS

RECORD LINKAGE METHODOLOGY

Chairperson: J.N.K. Rao, Carleton University

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

USING LARGE DATA BASES FOR RESEARCH ON SURGERY

LESLIE L. ROOS and NORALOU P. ROOS[1]

ABSTRACT

This paper reviews common characteristics of existing health care data bases, emphasizing the relationships among these characteristics and the questions researchers wish to ask. Although the uniform hospital-discharge data set is available for almost all Canadian provinces and in many American settings, additional information of critical importance for research design is not characteristic of many data sets. This information includes: coverage of an entire population, having a unique identifying number (or combination of identifiers), and presence of an enrollment file specifying when coverage begins and ends. Three levels of data are identified according to their comprehensiveness; each level is discussed in terms of the type of research designs which can be supported by the information available.

Level 3 data (only hospital discharge abstracts) can support analyses of utilization across medical market areas, as well as studies of length of stay and of inhospital mortality. Level 2 data (hospital discharge abstracts plus consistent individual identifiers) can be used for short-term studies of readmissions and complications after surgery. Such research on quality assurance and cost control can be performed in a timely manner to provide feedback to health care institutions. Level 1 data (hospital discharge abstracts, consistent individual identifiers, and an enrollment file) permit high quality longitudinal research; comparisons among alternative treatments and different hospitals can be made. Incident cases can be identified; analyses can be carried out across medical market areas on a per person basis. Other important uses of existing data bases — particularly Level 1 data — include the study of various aspects of physician behavior and the providing of information for clinical decision analyses.

Level 1 data are uniquely valuable in specifying possible biases and opportunities in less comprehensive data sets. This paper looks at case-mix measurement and hospital — versus population-based data, highlighting issues being addressed in ongoing research. By showing where less comprehensive information can be used with comparatively little loss in accuracy, such work may open the way for valuable studies in more jurisdictions by more investigators.

Two practical topics confronting researchers who wish to use administrative data are then addressed: the applicability of record linkage to produce more detailed data and the characteristics of software appropriate for working with existing data bases. Finally, the policy issue of how to improve surgical outcomes is discussed.

[1] Leslie L. Roos and Noralou P. Roos, Departments of Business Administration and Community Health Science, Faculties of Management and Medicine, University of Manitoba, Winnipeg, Manitoba, Canada. R3T 2N2.

1. INTRODUCTION

Health care data bases of varying scope and quality currently exist in a number of different settings; they are held by research groups, hospitals, insurers, and governmental agencies. Of particular interest are the data currently generated by health insurance systems in North America, Europe, Australia, and New Zealand. Large numbers of patient discharge abstracts filed by hospitals are being produced, and scholars are becoming increasingly aware of the research potential of these systems. Thus, the National Institute of Medicine (Committee for Evaluating Medical Technologies in Clinical Use, 1985) has recommended strengthening methods of evaluating medical practice using existing data bases. Feinleib [16] has stressed the number of investigations possible when rich, high-quality data bases can be made more accessible. Researchers, administrators, and practitioners are increasingly facing an information-rich environment, if they only knew what to do with the data! This paper addresses questions of: how should administrative data bases be constructed to facilitate research? How can existing data be best analyzed?

What would be the characteristics of an ideal data base for research on surgery? First of all, coverage of an entire population would permit looking at utilization from an epidemiologic perspective, attributing use to individuals according to their place of residence, no matter where the services are provided. Populations could then be compared in terms of how much of any given resource is used; such population-based data can be adjusted for age, sex, and other characteristics to facilitate comparisons.

In this ideal data set each person should be specified by a unique number or combination of identifiers. When each individual is so identified, usage can be cumulated for each person, wherever the care is received. This data base should capture all contacts with the health care system for each individual (with the unique identifier available to facilitate tracing). Among contacts which might be recorded would be all hospital care, both inpatient and outpatient, services in free standing surgery-centers, activities in physician offices, entry to nursing or personal home care , health care received in the home setting, and prescription drug use. Thus, an individual having surgery in one setting and being readmitted to a second institution will have both contacts captured by the system.

Finally, keeping track of when each individual's coverage begins and ends and why is very useful. Such a registry or enrollment file is necessary to deal with the issue: if an individual has no recorded contact with the health care system, was he resident within the jurisdiction and indeed had no contact? Alternatively, has he left or has he died?

Such a data base would facilitate not only comparative studies of utilization in different medical market areas, but also longitudinal research on utilization by groups of individuals (defined in many different ways). This type of nonintrusive follow-up has a number of advantages over follow-up by more labor-intensive methods [40]. Such follow-up can greatly aid research directed towards understanding the outcomes of various surgical interventions. Longitudinal designs are also essential for tracing the history of certain medical conditions, both those that result in hospitalization and those which do not.

Administrative data are particularly well-suited for studying major health outcomes: death, nursing home admission, and hospitalization. Other outcomes resulting in contacts with the health care system can be traced; in many cases the proportions of individuals enjoying "intervention-free survival" — survival without any contact with the health care system — can also be ascertained.

Because many data bases are maintained and updated for administrative purposes, the researcher, once having obtained access, may be able to analyze the data base for a relatively small marginal cost. Major costs would already have been underwritten by the

agency using the data for management purposes, and a variety of analyses would be possible. The number of years (of treatments or other events) can be increased either by bringing in more recent years of data or by going back farther in time. This facilitates long-term follow-up of individuals and the studies of rare events; more mundanely, adding extra years leads to a larger number of cases, increasing the statistical power of any analysis.

Data Base Limitations

Even such an "ideal" data base may have problems. First of all, the quality of the information which has been recorded must be assessed. Several Canadian provinces have data bases which have been used extensively for research purposes [3,75]. For example, in Manitoba, both the surgical procedures performed in hospitals and discrete billable items (even if not major events, including such things as pap smears) appear to be reliably captured in the claims system [42,47]. The accuracy of diagnostic data depends upon both the physicians and the clerks recording the diagnoses. American Medicare data appear to record procedures performed with fair accuracy, particularly if the "order of procedure" is ignored. Medicare data quality may have gone up since the introduction of the Prospective Payment System (PPS), but diagnostic information may not be as accurate as in the Manitoba files [12,74]. Medicare data also suffer from not including outpatient information in the hospital file.

Diagnoses on hospital records are likely to be more accurate than diagnoses on claims generated by physician's visits. In Manitoba, diagnoses are noted with a reasonable degree of accuracy and specificity in the hospital system, reflecting the professional training of medical records technicians. A comparison of diagnoses recorded on hospital records with those reported in the claims showed 95 percent correspondence in gallbladder disease, and 89 to 92 percent correspondence (representing identical diagnoses) in a study of acute myocardial infarction patients [41-42]. Manitoba currently records up to 16 diagnoses on every hospitalized patient using the internationally standardized ICD-9-CM system. Diagnoses from an ambulatory care system are useful, but at a much more generalized level. One approach which has proved fruitful in Manitoba has been to group diagnoses available from physician claims (for example, those reflecting contacts for gynecologic problems in a study of women undergoing hysterectomy, and those reflecting gallbladder disease — abdominal pain for a study of contacts before and after gallbladder surgery) rather than to attempt fine diagnostic distinctions [11-12].

Another problem relates to the detail of the coding conventions used. The widely used ICD-9-CM coding system does not distinguish procedures performed on the left side of the body from those done on the right side. This presents difficulties in assessing the results of orthopedic surgery; a second hip or knee replacement operation on someone who has already had one may mean either a reoperation on the same extremity or an operation on the other one [53].

No matter how much is provided in any data base, specific items desired for a given study may not be available. Additional information may — or may not — be contained in other sources which permit linkage to an existing data base. In particular, administrative data bases often do not include the performance of certain tests or x-rays if they are not billable; the results of tests are frequently not included. Information on medical treatment (such as drugs used) is typically not available, making it difficult to compare medical and surgical alternatives in the treatment of many conditions.

Finally, the individual rather than the researcher initiates contact with the system generating the data. Thus, a person who is ill, but has no contact with the health care system, does not produce a record which contains information on this particular episode of illness or chronic condition. The degree to which this affects health status measurement will vary both with certain system characteristics (such as the extent of insurance

coverage) and with individual characteristics (care seeking behavior). Given universal insurance, relatively few ill individuals do not have contact with the health care system [30]. In fact, over a four year period, 91 per cent of Manitoba adults contact a physician at least once.

The uniform hospital-discharge data set is likely to be available in many political jurisdictions. The following data seem commonly available on hospital claims collected in almost all Canadian provinces and in many American settings.

Such items include:

Date of birth
Sex
Place of Residence
Identifying Number (Individual or Family)

Other items for analysis include:

Discharge Diagnoses (several)
Procedures Performed in Hospital (several)
Hospital
Date of Admission
Date of Discharge
Discharge Code (Death, Another Hospital, Home, etc.)

Secondary items:

Admitting Physician Identifying Number
Physician Performing Each Procedure (Identifying Number)

As noted above, the following important information may — or may not — be available in a given data set:

- Coverage of an entire population
- **Unique** identifying number (or combination of identifiers)
- Enrollment or registry file specifying when coverage begins and ends

2. RESEARCH DESIGNS

Given variation in both the questions researchers wish to ask and the characteristics of the data to be accessed, Table 1 highlights the information required for different types of studies of cost control, quality of care, and treatment outcomes. The designs form a continuum from relatively simple to relatively complex, according to the data requirements imposed by each design. In this Table, level is identified on the basis of what can be done with the data. Thus, the most comprehensive information (Level 1 data) permits all the analyses possible using simpler Level 2 and Level 3 data, as well as supporting research which cannot be done using less complete data sets. The simpler the data requirements for a given design, the more political units (states, provinces, etc.) will have collected information so as to permit this particular type of research.

Table 1

Data Requirements and Types of Studies Using Hospital Data

Data Requirements	Types of Studies
Simple — Level 3 Need hospital discharge abstracts Length of Stay Small Area Analyses	In-hospital Mortality (Volume-Outcome Comparisons, Monitoring of Individual Hospitals)
Intermediate — Level 2 Need hospital discharge abstracts and consistent individual identifiers	Timely Longitudinal Research Short-Term Readmissions; Volume-Outcome Comparisons, Monitoring of Individual Hospitals for Quality Assurance and Cost Control
Comprehensive — Level 1 Need hospital discharge abstracts, consistent individual identifiers, and enrollment file Monitoring of Individual Hospitals, Choice of Treatment, Small Area Analysis by Person	Highest Quality Longitudinal Research Short-Term and Long-Term Outcome Studies; Identification of Incident Cases, Volume-Outcome Comparisons,

Designs with Simple Data Requirements (Level 3 Data)

The data requirements for some designs are relatively simple. If both a consistent individual identifier and an enrollment file are lacking, three types of cross-sectional studies are common. Such studies build on the information in the hospital discharge abstracts without trying to trace individuals. The data sets involved in such studies tend to be large, generally permitting comparisons across many hospitals. The hospital-focused studies of mortality and length of stay described below often are not population-based; thus, controls for case-mix are especially important.

Luft and his collaborators [27-28,59] have used data from hospitals subscribing to the Commission on Professional and Hospital Activities (CPHA) to demonstrate the importance of hospital volume as a variable affecting survival after complex surgical procedures. For many procedures, patients operated upon at hospitals performing a greater number of procedures do, on average, better than those operated upon at hospitals performing fewer procedures. Although Sloan et al. [61] present 1972-1981 data from 521 CPHA hospitals to argue that an adequate statistical basis for setting minimum volume standards does not presently exist, the bulk of evidence supports the volume-outcome relationship [26]. These cross-sectional analyses are limited to such outcomes as "mortality within the hospital stay in which surgery took place". Although better data will permit more long-term outcome studies, ongoing research suggests that, for many common surgical procedures, studies concentrating on inhospital mortality capture enough of the deaths which occur within three months of surgery to permit useful research.

Another type of study uses information on the hospital discharge abstract to classify patients into diagnosis-related groups (DRGs). Patients in a given group are assumed to have similar processes of hospital care and services provided [17]. Hospitals can be compared DRG by DRG, to determine which hospitals' patients have longer lengths of stay; an overall case-mix adjusted length of stay index can be computed for each hospital. American legislation mandating the use of DRGs as part of the Medicare prospective payment system has made such studies important for ongoing monitoring of hospital

utilization. Length of stay is being heavily used by hospital managers "as a surrogate measure for cost" [63]. The application of DRGs has generated considerable controversy; a number of efforts to improve on this method of controlling for case-mix are underway [10,20,64].

Small area analyses also have relatively simple data requirements, but they do require a population base. Hospital market areas are defined to include "any geographic subunit in which more of the people living there use that particular hospital than any other one (a 'plurality rule')" [6]. Such a population base provides a denominator (counting all individuals living in an area) and permits age and sex adjustment of the utilization experience of the population, thus removing one of the most important patient-related contributions to variation in utilization rates [55]. All health care utilization by persons in the area is counted, regardless of where it takes place.

Combining length of stay and small area analyses can highlight the most effective way to control utilization. Treatments are likely to differ in terms of whether the DRG approach (which emphasizes length of stay) or the small area approach (focusing on admission per capita) is most suitable for understanding the overall variation in total days spent in hospital [73]. Analyses from Manitoba dramatically illustrate this difference. Coronary artery bypass graft surgery is centralized in Manitoba with all procedures being done in two Winnipeg teaching hospitals. The physicians in western Manitoba refer relatively few patients to Winnipeg for coronary angiography and subsequent bypass surgery. Table 2 shows how variation in the bypass procedures per capita are more important than differences in lengths of stays for bypass patients in understanding regional differences in overall utilization. The right-hand column presents total days/utilization per 10,000 adults for the seven planning regions of Manitoba. This column is the product of the direct adjusted rate per 10,000 adults and the mean length of stay per admission for bypass surgery.

Table 2

Admission Rates, Length of Stay, and Total Days Per Capita for Coronary Artery Bypass Graft Surgery for Residents of Manitoba's Seven Regions (1979-84)

Region	Direct-Adjusted Admission Rate Per 10,000 Adults	Mean Length of Stay	Total Days Per 10,000 Adults
Central	2.93	22.32	65.42
Eastern	3.73	21.47	80.22
Interlake	4.06	18.80	76.36
Northern	5.15	20.67	106.38
Parkland	3.28	19.47	63.83
Western	2.40	20.47	49.05
Winnipeg	5.04	20.64	104.01

Mean annual admission rates, length of stay, and total days per 10,000 adults for coronary artery graft surgery are presented in this table.

From a policy perspective, both strategies — reducing admissions and shortening length of stay — can be effective in cost control. Both have played a role in the recent lower rates of increase in American hospital costs [57]. Such analyses are particularly important because admission rates for most medical and surgical hospitalizations vary among hospital market areas. Major differences are found even between Boston and New Haven, sites of major teaching centers linked to Harvard and Yale, respectively [72]. Since large differences in illness rates cannot adequately explain the differences in hospitalization rates seen in small area studies [39,55,71], this variation is increasingly the focus of attention [73]. Feedback has sometimes been successful in reducing rates which are very high [70]. These rate data are also essential for payment systems based on capitation, which are being vigorously promoted by the American Medicare program [23].

Designs with Intermediate Data Requirements (Level 2 Data)

Intermediate designs are suitable when consistent individual identifiers are included in the data set, but enrollment files are not readily available. Hospital claims can be sorted by date and identifying number to generate hospitalization histories for each individual. Without an enrollment file, when coverage began and ended for an individual can not be incorporated into the analysis. A focus on short-term outcomes — both morbidity and mortality — is therefore appropriate with these designs. Such short-term outcomes might include readmission for particular complications, readmission after time periods of interest (48 hours, 7 days, 6 weeks from original discharge), or mortality (when present on a subsequent hospital abstract). In many cases, if the time interval is fairly short (up to a year or so after surgery), loss to follow-up (due to migration or, in the U.S., due to change of insurer) can be ignored without significantly biasing the results. In the United States, this appears to be true for Medicare (but not Medicaid) recipients [24,71]. Designs which do not depend upon an enrollment file permit performing outcome research more efficiently and quickly than those which involve linking utilization and enrollment files. This may be important in providing rapid feedback for audit committees overseeing quality of care.

A computer algorithm using a combination of "time after surgery" and "diagnosis at time of readmission" for classifying hospital readmissions as postsurgical complications would be useful for monitoring purposes. Recent work with hysterectomy, cholecystectomy, and prostatectomy has shown the feasibility of developing computer algorithms for such classification; results produced using these algorithms closely corresponded with those from physician panels [38]. Such studies of short-term outcomes may involve printing out cases for more detailed investigation; a listing which identifies individuals having been readmitted to hospital due to certain diagnoses or during a particular period of time after surgery can significantly reduce the amount of paper which quality assurance committees must confront. Thus, the computer can act as a first screen; hospital records can then be pulled for further investigation.

Population-based (as compared with hospital-based) data help improve these intermediate designs. If individuals have an identifying number assigned by a single hospital, only readmissions to this hospital will be captured. If, however, unique identifying information is assigned by a larger, geographically-based insurance plan (such as some Canadian provincial plans and U.S. Medicare), data for a given patient can be aggregated across all providers reimbursed by the plan. This is especially important for capturing readmissions of rural patients who have had surgery in an urban hospital; many of their readmissions will take place in small hospitals back in their home areas. For example, 45 per cent of Manitoba rural patients readmitted for complications after cholecystectomy entered a hospital other than the one in which they had surgery [38]. Thus, when system-wide data are not available, record reviews will systematically underestimate postsurgical complication rates.

Designs with Comprehensive Data Requirements (Level 1 Data)

If both consistent individual identifiers and an enrollment file are present, loss to follow-up can be estimated. Longitudinal studies meeting various criteria desirable for high-quality cause-effect research can be conducted [21]. The ability to develop person-based longitudinal histories permits identifying the number of new (first-time) occurrences in a population. This identification of incident cases is important in generating a relatively homogeneous group for study; a second operation or recurrence of a condition can be distinguished from new events. Outcome studies can be markedly improved by such information; because a second bypass operation may well be riskier than the first such operation, analyzing the two separately is very helpful. In similar fashion, valve replacement surgery following a bypass operation can be distinguished from valve surgery with no prior patient history of major operations.

In a system with national health insurance, short-term studies incorporating an enrollment file will be only slightly more accurate than those without such a file [45]. When coverage is not universal (at least within a given age group, i.e. the elderly), checks against an enrollment file will be necessary [24]. Migration in and out of a given insurance plan may be large enough to substantially affect results.

Such long-term outcomes as reoperations, nursing home entry, and mortality are particularly suitable for study using comprehensive data bases. As noted elsewhere [56], information from an insurance system which provides complete coverage for a population regardless of where treated may uncover problems undetected in research based on data from a single hospital. As time passes, patients become increasingly likely to have received care in more than one hospital; single hospital data sets drift into increasing rates of error.

Research on both efficacy and effectiveness is greatly facilitated by system-wide data. Studies of efficacy, of results in the so-called "best" situation (generally a teaching hospital), have usually been reported in studies of technology assessment [5]. However, research on the efficacy of many procedures is lacking [34]. The relative paucity of solid outcome studies seems to support physician uncertainty as to choice of treatment which may underlie much of the data on small area variation [69]. Moreover, community hospital practices and medical care outcomes may differ from those publicized by researchers at academic centers. Effectiveness studies, studies which present outcome results from representative samples of all hospitals and all physicians, have seldom been done.

Recent studies of prostatectomy can serve as a model for longitudinal studies of a common surgical procedure [54,74]. Data from Maine and Manitoba were combined to follow men having a prostatectomy for up to eight years. Combining data from these two jurisdictions helped generate a larger number of cases. This in itself can be important because some outcomes, particularly postoperative mortality, are relatively infrequent and the numbers of providers is relatively small. Both morbidity (as judged through revisions, readmissions due to complications, and so forth) and postoperative mortality were measured. Overall, adverse outcomes (both deaths and nonfatal complications) were more frequent than had been noted in the literature. The data base facilitated the capture of admissions in a hospital other than that of surgery; this, and the longer period of follow-up, no doubt contributed to the higher rates of adverse outcomes found in this study compared with other studies (most of which were based in teaching hospitals). The wide range of outcomes among different hospitals reinforced the need for such studies of effectiveness. The adjusted odds ratios for mortality within three months of surgery ranged from 0.48 to 4.79 among individual hospitals. In one hospital the mortality rate was one-half as high as in the baseline hospital; in another institution the rate was almost five times as great. One particular subgroup, men resident in nursing homes, was found to have especially high mortality rates.

Cohort data can sometimes be used to compare the outcomes of two types of surgery or of medical versus surgical treatment. The prostatectomy research discussed above found statistically significant differences between the outcomes of open and transurethral procedures [44]. These differences are being investigated more fully in ongoing work. In another example, administrative data proved especially useful in accumulating a sufficient number of cases of infective endocarditis for analysis. Alternative treatments for this condition (which requires hospitalization on diagnosis) have been analyzed using a cohort design [1].

Other comparisons of treatment options have dealt with more common conditions. Manitoba cohort studies of tonsillectomy focused on patient variables as age, sex, and the number of preoperative episodes of respiratory illness in the treated and untreated groups. Since operated and unoperated siblings could be compared using the data base, family variables were also taken into account [37,43]. Sensitivity testing estimated the accuracy of the analysis of tonsillectomy outcomes. Postoperative outcomes of a subset of Manitoba patients — those matching the Pittsburgh randomized trial of tonsillectomy criteria of a preoperative history of seven episodes in the year before surgery — could be compared with outcomes of patients enrolled in the Pittsburgh trial. With patients equated for preoperative history, the Manitoba results (from the cohort comparisons) were quite similar to those from the Pittsburgh randomized trial. Manitoba and Pittsburgh patients showed similar magnitudes of improvement following tonsil surgery; the surgical patients experienced between one and one and a half fewer episodes of respiratory illness than did the non-surgical groups.

3. SUBSTANTIVE ISSUES

A number of issues are relevant across data sets and research designs. First of all, in addition to focusing on the patient, administrative data bases can be fruitfully used to examine various aspects of physician behavior. This is important, since physicians are responsible for decisions that govern the way that as much as 90 per cent of each health care dollar is spent [14]. Although some research can be carried out using cross-sectional data, most work on physician behavior will depend upon comprehensive Level 1 data. Doctors enter and leave geographic areas and insurance plans; data on physician enrollment are a necessity for many studies.

Existing data bases should have an increasing impact on the burgeoning field of clinical decision analysis [35]. Abrams et al. [1] note that "decision analysis has helped to clarify the structure of some management controversies, balancing the risks and benefits of different strategies in a quantitative fashion". Decision analytic techniques, estimating the probabilities associated with alternative treatment choices, call out for better information on outcomes. Such data are sketchy for many treatments, even those which have been in general use for a number of years. Combining decision-analysis with claims data bases seems "useful as an alternative or precursor to randomized trials", especially when the difficulties of performing such trials are great [1]. Longitudinal data — Level 1 or possibly Level 2 data — are required for such studies.

Relatively complete data sets are invaluable in estimating the limitations of less complete data bases. Analyses of comprehensive Level 1 data can provide a practical "gold standard" against which other data can be compared. This permits estimating the practicality of a particular study in a jurisdiction with less comprehensive data. Eventually, feeding back results to those responsible may lead to changes in data collection. Both the section on case-mix and the following one comparing hospital-based and population-based approaches use the more complete data sets to highlight problems with the less complete.

Physician Behavior

Small area and hospital analyses imply a concern about physician practice patterns [13]. If physician identifiers (admitting physician, physician performing surgery) are available, physicians can be studied with the same techniques used for analyzing hospitals. Both volume-outcome relationships and the monitoring of individual physicians are legitimate areas of research and policy interest using administrative data bases [26,39].

Physician referral behavior also deserves more study. In several Canadian provinces and American states, more complicated cases are routinely referred from rural areas and small towns to urban tertiary centers distant from the patient's place of residence [25]. As discussed earlier, some differences in surgical rates across areas have been shown to be due to differences in the rate of referral to tertiary centers [52]. Such findings supported the hypothesis that having a small number of centrally-located physicians perform a particular procedure may contribute to the maintenance of regional variation; Manitoba data indicate that differences in referral behavior may continue over long periods of time. Because of the previously cited findings "of an inverse relation between the number of patients treated with specific diagnoses or procedures in a hospital and subsequent adverse outcomes" [29], policies directed toward centralization must deal with the possibility of adversely affecting equity of access.

A physician's hospitalization practice style is a third research topic facilitated by existing data bases. Since "only a small proportion of hospitalizations fit a model based on medical need" [67], researchers typically assume that significant differences in physician practice patterns are the primary reason for small area variations in both surgical and nonsurgical hospitalization usage rates [18]. Although these assumptions about physician practice style and patient need have seldom been tested directly, they have important implications for health planners. If need can be met with more cost-effective medicine, then health care spending can be restricted without rationing [67].

The comprehensive Manitoba data have facilitated developing an index of physician hospitalization style based on:

1. identifying physicians' primary patients
2. calculating the expected rate of hospitalization for these primary patients over a two year period controlling for differences in health status
3. comparing each physician's expected rate of hospitalizing his patients with the observed (actual) hospitalization patterns of his primary patients [52].

This index is being used to assess the influence of physician style on the utilization of hospitals; to test the stability of practice style over a long-term period; and to look at the reasons for, and outcomes of, hospitalization for patients of "hospitalization-prone" and "nonhospitalization-prone" physicians. Both overall hospitalization and surgical hospitalizations can be studied using this methodology.

Clinical Decision Analysis

The field of clinical decision analysis can benefit greatly from the information in existing data bases. For example, a clinical decision analysis of treatment alternatives for infectious endocarditis has provided insights into how to use retrospective chart reviews from a single hospital in tandem with a large claims data base [1]. Probabilities derived from the two data sources were remarkably similar, given the relatively low N for both the single hospital (16 cases) and provincial hospital claims data (127 cases). The use of the Manitoba data base, aggregating cases over a five year (1979-84) period, made feasible the analysis of survival and the comparisons of alternative treatments.

Researchers working in the field of clinical decision analysis have developed sophisticated methods of sensitivity testing. Given that the individuals in different "arms" of a cohort study (i.e., those having a particular operation at different hospitals, those receiving different treatments) are not randomly assigned, such sensitivity testing is very useful in establishing a particular difference in outcomes in the face of possible measurement error.

Clinical decision analysis also needs to incorporate the previously discussed research showing major differences among providers in case-mix adjusted outcomes following surgery. The probabilities for different outcomes clearly vary among hospitals and surgeons; data on such variation are now available for about half a dozen surgical procedures [26]. Decision analyses might present "high quality treatment" and "low quality treatment" decision trees. While such results might be controversial, they may accurately reflect the real world.

Case-Mix Adjustment

Controlling for possible differences in patient characteristics is critical for comparing treatments and providers. Deciding what types of data are necessary for adequate controls to study length of stay, death, disability, and rehospitalization is not simple. As Jencks and Dobson [22] have pointed out, "severity adjustments that predict outcome are not identical to those that predict cost". Research on this topic is very timely because decisions about expenditures for additional data collection are being made. Thus, Pennsylvania Health Care Cost Containment Council has recently required extra data (collected by MedisGroups) to be added to the hospital discharge abstract to facilitate more detailed risk adjustment for quality assessment [22]. The necessity of such expensive prospective data for case-mix adjustment needs to be established.

If cross-sectional data can provide enough information for case-mix controls, then large-scale studies of inhospital mortality following surgery will be relatively easy to conduct. Researchers have to establish empirically the extent to which information from claims at each level:

a) Level 3 - cross-sectional

b) Level 2 - intermediate (longitudinal without an enrollment file)

c) Level 1 - comprehensive (longitudinal with an enrollment file) is sufficient for case-mix controls. Different outcomes may well require different types of controls.

As noted earlier, considerable effort has been — and is being — exerted to try to cost-effectively improve measurement of case-mix using cross-sectional data. Age, sex, and comorbidity data from the surgical hospitalization have been used to control for interhospital differences in Luft and his collaborators' research [28,59]. Using information on the claim for each hospital stay, DRGs often control for age, sex, and (sometimes) comorbidity. Jencks and Dobson [22] feel "no available measure of severity of illness" would markedly improve the accuracy of Medicare payments "if used to supplement or replace the system of diagnosis-related groups".

When consistent individual identifiers are present, additional covariates can be obtained from data available from the period preceding the index hospitalization (i.e. the hospitalization of surgery). A methodology for conducting outcomes research using covariates from prior claims is being developed [44,72]. Such covariates are not subject to the potential biases of covariates derived from cross-sectional data. For example, in a cross-sectional study the diagnosis of "myocardial infarction" may represent a pre-existing condition or an event which occurred after surgery but during the surgical hospitalization [4].

The availability of an enrollment file helps insure that individuals whose coverage began just before surgery will not be confused with those having more complete prior histories. Such a file is desirable; its necessity depends on the population being studied. For Medicaid recipients in two states, only between 54 per cent (Georgia) and 63 per cent (Michigan) were enrolled for a full six months before and after hospital admission for one of eight selected surgical procedures [24]. When a system of universal health insurance is in place, the results of considerable sensitivity testing show an enrollment file is not essential when the presurgical period is short (up to two years) and migration is comparatively low [45]. Thus, longitudinal analyses of follow-up using Medicare data suggest case-mix adjustment without checking an enrollment file is unlikely to present major problems.

The appropriate strategy for developing measures of case-mix adjustment needs attention. Because of the relatively few adverse outcomes, large data sets are highly desirable for case-mix adjustment. Some strategies — such as the "condition/treatment" model proposed by Wennberg [68] — depend on using a number of dichotomous, independent variables (for example, presence or absence of prior cancer) as separate predictors; however, for most surgical and medical treatments, comparatively few cases of specific conditions (for example, presence of cancer) are available. Without a very large number of cases, combining some of the independent variables into an index of comorbidity or illness severity may be essential for case-mix adjustment.

Mosteller et al. [31] note that "the potential gain from measurement offers one reason for developing scales of measurement". Indices and scales offer the promise of both increased reliability and use of a metric rather than a dichotomy; consequently, the required sample size can be smaller. Charlson et al. [7] have developed and validated a method of classifying comorbidity to estimate risk of death. Because the Charlson Index was constructed from hospital medical records, using either cross-sectional claims data or claims recorded during the presurgical period to generate the index should also be possible. For several common surgical procedures, hospital claims alone provide almost as good case-mix adjustment as does linking information from hospital claims with that taken from physician visits or from more costly survey information [30,56]. Ongoing work suggests that using **both** claims prior to the surgical hospitalization and those generated by this hospitalization provides the best case-mix adjustment.

In the previously-cited prostatectomy research, case-mix was measured using all available hospital claims — both those in the six months prior to surgery and those for the individual at the time of surgery [74]. Relevant items from the six months before prostatectomy included: hospitalized with cancer (except prostate), hospitalized with cardiovascular diagnoses, and resident of a nursing home. Associated diagnoses at the time of surgery included: cancer of prostate (not previously diagnosed), cancer (except prostatic), cardiovascular diseases, and other associated diagnoses (more than one diagnosis given). In addition to comparing hospitals, this study has also compared open versus transurethral prostatectomy with regards to post-operative mortality rates. Since transurethral procedures are generally thought appropriate for patients too ill to undergo open procedures, developing adequate covariates to adjust across the two groups was clearly key.

To address this problem, claims data were combined with data collected prospectively from one Manitoba hospital [8-9]. In this hospital, every surgical patient was interviewed by a nurse collecting information about preoperative drug use (the number of drugs used as well as which specific drugs-digitalis, etc.) and the number of preoperative conditions which affected the patient (including obesity, respiratory problems, etc.). In addition, the anesthesiologists rated each patient using the American Society of Anesthesiologists' Physical Status Classification (ASAPS) [33]. This five point scale runs from 1 — healthy normal patients to 5 — those not expected to survive surgery.

Two logistic regression models were developed for predicting the occurrence of death in the period following surgery. One model included variables derived from administrative data alone; the second model added prospectively collected data (ASAPS, etc.) to the variables available from administrative data. Two of the variables collected prospectively (ASAPS and being on digitalis) entered the equation with significant coefficients, replacing four of the six variables derived from claims data. The model which included prospective data had only slightly better predictive power; both models fit the data reasonably well. Moreover, the coefficient indicating the association of type of surgery with postoperative mortality was identical in the two models, suggesting that administrative data provided very useful controls for case severity [49]. These analyses are being conducted across a number of surgical procedures; if the prostatectomy findings are generalizable, they will provide strong evidence that claims data can do almost as well in providing measures of comorbidity as can measures based on expensive primary data collection.

When are administrative data "good enough" for distinguishing the better of two treatments, for testing hypotheses about the relationship between surgical volume and treatment outcomes, and for identifying hospitals with particularly poor (or especially good) outcomes? In studying the link between volume and outcomes, one needs only to show that the average case-mix in high-volume hospitals is no different than that in their low-volume counterparts. Statistical adjustments are much more likely to be necessary when comparing treatment alternatives and monitoring individual hospitals. Both because the true differences in case-mix are not known and because there is a limit as to how much case-mix can be adjusted at present levels of knowledge, the results of any technique must be compared with a standard of what can be done given the **best available** techniques.

As seen above, improving upon what can be done using administrative data is often difficult. Given the costs of gathering data on risk factors prospectively, extensive data collection seems hard to justify by the available evidence. Prospective data collection for a small number of hospitals may be helpful in gathering risk factor data **plus** information on postoperative condition or on possible causal factors associated with particular outcomes.

Hospital-Based versus Population-Based Data

Analyses of outcomes of care are facilitated by population-based data on all deaths and all readmissions, not just events occurring in the hospital of surgery. On the other hand, data from the hospital of surgery may be more widely available and easier to analyze, facilitating more timely feedback and monitoring of quality of care. Analyses which are hospital-centered (either because the investigators are based at a single hospital or because the individuals being studied are identified by a number specific to one hospital) may miss important events in following up the patient population.

Such work is particularly timely because of the concern that Medicare's prospective payment system might encourage such short stays that quality of care would be adversely affected. If this were so, readmissions and mortality which might be avoidable could be missed or assigned to the wrong hospital by a monitoring system. Having comprehensive data facilitates improvement of outcome analyses. Cross-sectional, hospital-based analyses which do not depend on individual identifiers or enrollment files can be compared with longitudinal, population-based analyses capturing all readmissions and all deaths within a specified period. Thus, the more comprehensive data (typically available from fewer jurisdictions) can indicate for which procedures and for which types of patients inhospital mortality is an adequate measure of mortality in the period after surgery and for which it is not. Preliminary analyses suggest that for certain procedures (coronary artery bypass graft surgery, valve replacement surgery) deaths during the surgical hospital

stay represent a very high proportion of deaths in the three months after surgery. For other procedures and conditions (prostatectomy, hip fracture), this is not the case.

4. PRACTICAL CONSIDERATIONS

Two practical issues confronting researchers wishing to use administrative data are: how can I get better (more comprehensive, more detailed) data? and what kind of software can help efficiently analyze these data? This section discusses the application of record linkage techniques to produce better data and the characteristics of software for working with existing data bases.

Record Linkage

Record linkage has often been used to add mortality information to data on workers' exposure to possible health hazards in research on occupational health [62]. Other linkages with mortality data have involved disease cohorts, life style/risk factors, clinical trials, and general population cohorts. When enough identifying variables are available, record linkage permits generating Level 1 data for studies of survival. However, record linkage can also be used more generally to merge administrative data with other studies [66]. The statistical techniques on which record linkage rests have been treated extensively elsewhere. This discussion notes how record linkage as a way of thinking may suggest opportunities otherwise overlooked. Given a unique identifier (or a set of other identifiers), bringing in additional research (based on surveys, hospital record reviews, and so forth) to augment the information on each individual is often possible. For example, the Manitoba Longitudinal Study on Aging has retrospectively linked survey information and claims to provide a fuller picture of the relationship between self-reported health status and utilization. Cancer registry data and vital statistics records have also been combined with Manitoba Health Services Commission registry and claims information to improve the quality of data available in each file.

Looking for record linkage possibilities is a useful mental exercise when dealing with administrative data bases. Perhaps the researcher should look first at what would be desirable, and worry about what is practical later. This approach has led to several different linkages in our Manitoba work on common surgical procedures:

1. linkage of hospital claims with primary data on anesthesia and its outcomes to produce a very rich data set
2. linkage of hospital claims with physician claims to verify fact of surgery and to add operation date to the hospital claims
3. linkage of the enrollment file with Vital Statistics information to verify deaths and provide cause of death information

Although the linkage techniques used differed in each example, the linked files all proved helpful. Such linkages permit analysis at the individual level and dramatically increase the amount and quality of data.

Such data can be generated elsewhere. If names and/or identifying numbers are not available on two potentially linkable files, investigators should look for four or five similar variables on the two data sets. Those may be all that are necessary to both link the files and use proven methods for assessing quality of the matches [47,66].

Software

The above discussion has concentrated on conceptual building blocks for working with administrative data: research design, data base quality and comprehensiveness, definition

of outcomes, and case-mix measurement. On the practical side, appropriate software represents one step towards making the approaches described here available to a larger audience. A system to facilitate small area analysis, longitudinal studies, record linkage, and so forth is needed to efficiently use existing data bases. When a number of different analysts want to work with information formatted in various ways, easy to use software is particularly important.

In Canada the provincial health care bodies vary considerably in the amount of information collected and the format used for data management. Although items on the uniform hospital discharge data set seem available for each province, analysis is characteristically carried out at provincial rather than national levels. With considerable effort, Statistics Canada is able to put together a minimum hospital data set for the country as a whole. In similar fashion, the collection of American Medicare Part B data is decentralized to at least the state level, with an accompanying variety of formats. Software to improve their "inhouse" analytical capability might interest such groups as the provincial Colleges of Physicians and Surgeons in Canada, Peer Review Organizations (PROs) in the United States, insurers of various types, and hospital associations across North America.

In addition to the general capabilities outlined above, such a system should also be versatile, user-friendly, and reasonably economical of computer time. Speed of system development, ease of modification, and operation on a variety of machines are important factors. In the authors' experience, the widely adopted SAS language has proved ideal for development of a management information system for health analysts. SAS features flexibility in combining several variables (such as diagnoses and procedures) and good data subsetting capabilities for producing an abbreviated file containing selected claims and variables. SAS features facilitate working across several data bases in the same run. Increasing computer power makes the greater running time associated with using such high level languages of less concern [19].

SAS' macro processor and accompanying macro language were used to develop procedures to perform the functions outlined above. This macro processor "provides a way to store and retrieve SAS jobs that must be tailored to changing details". When appropriate, commands are common across the programs and the syntax in the modules corresponds to that used by SAS. Regular SAS features are used to manipulate and analyze the data. In this way, a coherent software system can be developed rapidly by building on existing fourth generation languages.

5. DISCUSSION

This paper has reviewed approaches to the information-rich environment available to researchers, policy makers, and managers. Research on cost-control and quality of care can have an obvious impact on the amount and distribution of resources for health care. Issues in identifying institutions with outcomes which should be carefully reviewed have been highlighted. The comparisons of the outcomes of different treatment alternatives are directly applicable to the emerging fields of clinical epidemiology and medical decision-making.

Existing administrative data bases are important for monitoring outcomes of many procedures, given both the North American "explosion in surgical utilization among the elderly" [65] and the great concern over health care costs. Specifying particular variables as predictors of postsurgical outcomes can help the physician decide as to the appropriate treatment alternative. Thus, the reporting of relatively high mortality following prostatectomies on nursing home patients resulted in a reduction of such operations by Maine urologists [11].

Given the limited population entering most clinical trials, cohort studies using claims data bases may be the major source of outcome data for operations performed on some age groups. For example, only claims data can provide ongoing information on mortality, readmissions, and "intervention-free" survival following coronary artery bypass graft surgery among the elderly. This procedure has grown rapidly among patients over 65 years of age, with some longitudinal series but no accompanying information from randomized clinical trials [2,46].

What will it take to improve outcomes of surgery? Although the U.S. and Canadian experiences in cost control have differed, health care expenditures have risen in both countries since the early 1970s. Such increases have occurred in delivery of care in specific institutions [58,60] and at a systems level [15,36]. It is not cost of health care per se, but cost in relation to quality of care and outcomes which is most important) [58]. However, data relating specific cost increases to particular changes in quality of care are almost uniformly unavailable. Increases in direct hospital expenditures over the past decade should have improved the quality of care. More intensive monitoring of surgical patients, more aggressive postoperative care, and more highly trained nursing staff are all developments which have occurred across North America. The major rationale for these changes has been to improve quality. Over the past ten years, Manitoba also experienced a centralization of high risk procedures and a limitation of the surgical privileges of nonspecialists.

However, research has shown little or no improvement in the outcomes associated with three common surgical procedures — hysterectomy, cholecystectomy, and prostatectomy -performed over a ten year period in Manitoba [44]. Three surgical cohorts (all those undergoing surgery in 1972-73, 1977-78, and 1982-83) were examined without finding significant decreases in the postoperative mortality rate, in the rate of readmission to hospital in the immediate postoperative period, or in readmission to hospital for adverse surgical outcomes in the 15 months following surgery. These data suggest "flat of the curve" medicine [32]. In the early 1970s, the organization and delivery of care for these three procedures was of such a quality that improving surgical outcomes proved difficult, despite ever increasing expenditures on the hospital sector.

Careful monitoring of, and feedback to, institutions with poorer-than-expected outcomes seems likely to have a much greater impact on quality than the incremental system-wide increases in expenditures which are the norm. Given reasonably high overall quality, achieving even small improvements are difficult. An emphasis on monitoring and feedback directs efforts toward the hospitals where improvement is more feasible. Where cost control is the issue, the techniques discussed here make it possible to focus on particular areas or particular hospitals rather than cutting across the board.

The maintenance, analysis, and improvement of existing data bases represents a cost-effective way to better understand issues of access to and quality of care. The technology and data are available; now funding to conduct the necessary research and the will to put findings into practice are necessary.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the help of the Manitoba Health Services Commission. This research was supported by National Health Research and Development Project No. 6607-1197-44 and by Career Scientist Awards No. 6607-1314-48 (to L.L. Ross) and 6607-1101-22 (to N.P. Ross). Interpretations and viewpoints contained in this paper are the authors' own and do not necessarily represent the opinion of either the Manitoba Health Services Commission or Health and Welfare Canada. The authors also wish to thank Sandra Sharp and Kerry Meagher for their help with the manuscript.

REFERENCES

[1] Abrams, H.B., Detsky, A.S., Roos, L.L., et al.: Is there a role for surgery in the acute management of infective endocarditis? A decision analysis and medical data base approach, Med. Decis. Making, 1988. (Forthcoming.)

[2] Anderson, G.M., and Lomas, J.: Monitoring the diffusion of technology: coronary artery bypass graft surgery in Ontario, Am. J. Public Health, 1988. (Forthcoming.)

[3] Barer, M.L., Evans, R.G., Hertzman, C., et al.: Aging and health care utilization: new evidence on old fallacies, Soc. Sci. Med. 24:851, 1987.

[4] Blumberg, M.S.: Risk adjusting health care outcomes: a methodologic review, Med. Care 43:351, 1986.

[5] Brook, R.H., and Lohr, K.N.: Efficacy, effectiveness, variations, and quality: boundary-crossing research, Med. Care 23:710, 1985.

[6] Caper, P.: Variations in medical practice: implications for health policy. Health Affairs 3(2):110, 1984.

[7] Charlson, M.E., Pompei, P., Ales, K.L., et al.: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation, J. Chron. Dis. 40:373, 1987.

[8] Cohen, M.M., and Duncan, P.G.: Physical status score and trends in anesthetic complications, J. Clin. Epidemiol. 41:83, 1988.

[9] Cohen, M.M., Duncan, P.G., Pope, W.D.B., et al.: A survey of 112,000 anesthetics at one teaching hospital (1975-83), Can. Anaesth. Soc. J. 33:22, 1986.

[10] Cretin, S., and Worthman, L.G.: Alternative systems for case mix classification in health care financing (Rand R-3457-HCFA), Santa Monica, CA, 1986, Rand Corporation.

[11] Davis, H.: Was surgery needed? The Baltimore Sun, Apr. 6, 1986.

[12] Demlo, L.K., and Campbell, P.M.: Improving hospital discharge data: lessons from the National Hospital Discharge Survey, Med. Care 19:1030, 1981.

[13] Egdahl, R.H.: Ways for surgeons to increase the efficiency of their use of hospitals, N. Engl. J. Med. 309:1184, 1983.

[14] Eisenberg, J.M.: Physician utilization: the state of research about physicians' practice patterns, Med. Care 23:461, 1985.

[15] Evans, R.G.: Finding the levers, finding the courage: what have we learned about cost containment in North America? J. Health, Politics, Policy, & Law 11:585, 1987.

[16] Feinleib, M.: Data bases, data banks and data dredging: the agony and the ecstasy, J. Chron. Dis. 37:783, 1984.

[17] Fetter, R.B., Shin, Y., Freeman, J.L., et al.: Case mix definition by diagnosis-related groups, Med. Care 18:1 (suppl), 1980.

[18] Griffith, J.R., Restuccia, J.D., Tedeschi, P.J., et al.: Measuring community hospital services in Michigan, Health Serv. Res. 16:135, 1981.

[19] Harel, E.C., and McLean, E.R.: The effects of using a non-procedural computer language on programmer productivity, MIS Quarterly 9:109, 1985.

[20] Hornbrook, M.C.: Techniques for assessing hospital case mix, Annu. Rev. Public Health 6:295, 1985.

[21] Horwitz, R.I.: The experimental paradigm and observational studies of cause-effect relationships in clinical medicine, J. Chron. Dis. 40:91, 1987.

[22] Jencks, S.F., and Dobson, A.: Refining case-mix adjustment: the research evidence, N. Engl. J. Med. 317:679, 1987.

[23] Jencks, S.F., Dobson, A.: Strategies for reforming Medicare's physician payments: physician diagnosis-related groups and other approaches, N. Engl. J. Med. 312:1492, 1985.

[24] Klingman, D., Pine, P., Simon, J.: Outcomes of surgery among Medicaid recipients in Georgia and Michigan, 1981-1982, 1987 (submitted for publication).

[25] Luft, H.S.: Regionalization in medical care, Am. J. Public Health 75:125, 1985.

[26] Luft, H.S., Garnick, D.W., Mark, D., et al.: Evaluating research on the use of volume of services performed in hospitals as an indicator of quality, Washington, DC, 1987, Office of Technology Assessment, Congress of the United States. (Draft.)

[27] Luft, H.S., Bunker, J.P., Enthoven, A.C.: Should operations be regionalized? The empirical relation between surgical volume and mortality. N. Engl. J. Med. 301:1364, 1979.

[28] Luft, H.S., and Hunt, S.S.: Evaluating individual hospital quality through outcome statistics, J.A.M.A. 255:2780, 1986.

[29] Maerki, S.C., Luft, H.S., Hunt, S.S.: Selecting categories of patients for regionalization: implications of the relationship between volume and outcome, Med. Care 24:148, 1986.

[30] Mossey, J.M., and Roos, L.L.: Using insurance claims to measure health status: the illness scale, J. Chron. Dis. 40:41S (suppl), 1987.

[31] Mosteller, F., Gilbert, J.P., McPeek, B.: Reporting standards and research strategies for controlled trials, Con. Clin. Trials 1:37, 1980.

[32] Neuhauser, D.: Cost-effective clinical decision-making implications for the delivery of health services. In Bunker, J.P., Barnes, B.A., Mosteller, F., editors: Costs, risks, and benefits of surgery, New York, 1977, Oxford University Press.

[33] Owens, W.D., Felts, J.A., Spitznagel, E.L., Jr.: ASA physical classifications: a study of consistency of ratings, Anesthesiology 49:239, 1978.

[34] Patricelli, R.E.: Employers as managers of risk, cost, and quality, Health Affairs 6(3):75, 1987.

[35] Pauker, S.G., and Kassirer, J.P.: Decision analysis, N. Engl. J. Med. 316:250, 1987.

[36] Reinhardt, U.E.: Resource allocation in health care: the allocation of lifestyles to providers, Milbank Mem. Fund Q. 65:153, 1987.

[37] Roos, L.L.: Alternative designs to study outcomes: the tonsillectomy case, Med. Care 17:1069, 1979.

[38] Roos, L.L., Cageorge, S.M., Austen, E., et al.: Using computers to identify complications after surgery, Am. J. Public Health 75:1288, 1985.

[39] Roos, L.L., Cageorge, S.M., Roos, N.P., et al.: Centralization, certification, and monitoring: readmissions and complications after surgery, Med. Care 24:1044, 1986.

[40] Roos, L.L., Nicol, J.P., Cageorge, S.M.: Using administrative data for longitudinal research: comparisons with primary data collection, J. Chron. Dis. 40:41, 1987.

[41] Roos, L.L., Nicol, J.P., Johnson, C., et al.: Using administrative data banks for research and evaluation: a case study, Eval. Q. 3:236, 1979.

[42] Roos, L.L., Roos, N.P., Cageorge, S.M., et al.: How good are the data? Reliability of one health care data bank, Med. Care 20:266, 1982.

[43] Roos, L.L., Roos, N.P., Henteleff, P.D.: Assessing the impact of tonsillectomies, Med. Care 16:502, 1978.

[44] Roos, L.L., Roos, N.P., Sharp, S.M.: Monitoring adverse out-comes of surgery using administrative data, Health Care Fin. Rev. 7:5 (suppl), 1987.

[45] Roos, L.L., and Sharp, S.M.: Becoming more efficient at outcomes research, Intl. J. Tech. Asses. Health Care, 1988 (Forthcoming.)

[46] Roos, L.L., Sharp, S.M.: Innovation, centralization, and growth: coronary artery bypass graft surgery in Manitoba, 1987. (Submitted for publication.)

[47] Roos, L.L., Sharp, S.M., Wajda, A.: Assessing data quality: a computerized approach, Soc. Sci. Med., 1988. (Forthcoming.)

[48] Roos, L.L., Wajda, A., Nicol, J.P.: The art and science of record linkage: methods that work with few identifiers, Comput. Biol. Med. 16:45, 1986.

[49] Roos, N.P.: Differential use of outpatient surgery by hospital physicians: what are the potential savings? Can. Med. Assoc. J. 1988. (Forthcoming.)

[50] Roos, N.P.: Hysterectomies in one Canadian province. A new look at risks and benefits, Am. J. Public Health 74:39, 1984.

[51] Roos, N.P., Danzinger, R.G.: Assessing surgical risks in a population: patient histories before and after cholecystectomy, Soc. Sci. Med. 22:571, 1986.

[52] Roos, N.P., Flowerdew, G., Wajda, A., et al.: Variations in physicians' hospitalization practices: a population-based study in Manitoba, Canada, Am. J. Public Health 76:45, 1986.

[53] Roos, N.P., and Lyttle, D.: Hip arthroplasty surgery in Manitoba: 1973-1978, Clin. Orthop. (199):248, 1985.

[54] Roos, N.P., and Ramsey, E.: A population-based study of prostatectomy: long term outcomes associated with differing surgical approaches, J. Urol. 137:1184, 1987.

[55] Roos, N.P., and Roos, L.L.: High and low surgical rates: risk factors for area residents, Am. J. Public Health 71:591, 1981.

[56] Roos, N.P., Roos, L.L., Mossey, J.M., et al.: Using administrative data to predict important health outcomes: entry to hospital, nursing home and death, Med. Care 1988. (Forthcoming.)

[57] Schwartz, W.B.: The inevitable failure of current cost-containment strategies: why they can provide only temporary relief, J.A.M.A. 257:220, 1987.

[58] Scitovsky, A.A.: Changes in the costs of treatment of selected illnesses, 1971-1981, Med. Care 23:1345, 1985.

[59] Showstack, J.A., Rosenfeld, K.E., Garnick, D.W., et al: Association of volume with outcome of coronary artery bypass graft surgery: scheduled vs. nonscheduled operations, J.A.M.A. 257:785, 1987.

[60] Showstack, J.A., Stone, M.H., Schroeder, S.A.: The role of changing clinical practices in the rising costs of hospital care, N. Engl. J. Med. 313:1201, 1985.

[61] Sloan, F.A., Perrin, J.M., Valvona, J.: In-hospital mortality of surgical patients: Is there an empiric basic for standard setting? Surgery 99:446, 1986.

[62] Smith, M.E.: Record linkage: organizing the facts together. In Bennett, B.M., Trute, B., editors: Mental health information systems: problems and prospects, New York, 1984, Edwin Mellen Press.

[63] Stern, R.S., Epstein, A.M.: Institutional responses to prospective payment based on diagnosis-related groups: implications for cost, quality, and access, N. Engl. J. Med. 312:621, 1985.

[64] Thomas, J.W., Ashcraft, M.L.S., Zimmerman, J.: An evaluation of alternative severity of illness measures for use by university hospitals, Ann Arbor, Mich, 1986, Department of Health Services Management and Policy, School of Public Health, University of Michigan.

[65] Valvona, J., Sloan, F.: Rising rates of surgery among the elderly, Health Affairs 4(3):108, 1985.

[66] Wajda, A., Roos, L.L.: Simplifying record linkage: software and strategy, Comput. Biol. Med. 17:239, 1987.

[67] Wennberg, J.E.: Commentary: on patient need, equity, supplier-induced demand and the need to assess the outcome of common medical practices, Med. Care 23:512, 1985.

[68] Wennberg, J.E.: Commentary: using claims to measure health status, J. Chron. Dis. (suppl) 40:51S, 1987.

[69] Wennberg, J.E.: Which rate is right? N. Engl. J. Med. 310:310, 1986.

[70] Wennberg, J.E., Blowers, L., Parker, R., et al.: Changes in tonsillectomy rates associated with feedback and review, Pediatrics 59:821, 1977.

[71] Wennberg, J.E., Fowler, F.J.: A test of consumer contribution to small area variations in health care delivery, J. Maine Med. Assoc. 68:275, 1977.

[72] Wennberg, J.E., Freeman, J., Culp, W.J.: Are hospital services rationed in New Haven or over-utilised in Boston? Lancet, May 23:1185, 1987.

[73] Wennberg, J.E., McPherson, K., Caper, P.: Will payment based on diagnosis-related groups control hospital costs? N. Engl. J. Med. 311:295, 1984.

[74] Wennberg, J.E., Roos, N.P., Sola, L., et al.: Use of claims data systems to evaluate health care outcomes: mortality and reoperation following prostatectomy, J.A.M.A. 257:933, 1987.

[75] West, R., Sherman, G.J., Downey, W.: A record linkage study of valproate and malformations in Saskatchewan, Can. J. Public Health 76:226, 1985.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

MISSING IDENTIFIERS AND THE ACCURACY OF INDIVIDUAL FOLLOW-UP

MARTHA E. FAIR and PIERRE LALONDE[1]

ABSTRACT

Wherever forms are being filled out, the need to record certain personal identifiers is constantly open to review by those who wish to minimize the labour and the possible unfavourable responses from the informants. The extent to which linkage error rates are influenced by the availability or non-availability of various identifiers is examined. This test is done by temporarily suppressing certain of the more important identifiers (e.g. the forenames in full, date of birth, and mother's maiden surname) contained in the search record singly and in groups. The searches are then repeated, with certain items present and with certain of them removed or not used. Data are presented from a mortality study of Ontario miners. The full date of birth is of special importance in this connection, and the full given name(s) come next. These data have important implications for the design of survey questionnaires, and for investigating the feasibility and anticipated accuracy of linkage projects. Where the readily available identifiers are limited, a methodology is examined for supplementing them. The strategy utilizes a supplementary file (e.g. Social Insurance Number file) to facilitate and improve the quality of the linkage.

1. INTRODUCTION

The purpose of this paper is to describe results of tests we completed to provide some quantitative data on the extent to which the accuracy of probabilistic record linkage of two administrative data files, such as a file of Ontario miners with the Canadian Mortality Data Base (CMDB), is dependent upon the presence or absence of various personal identifiers or groups of identifiers. This study relates to work at Statistics Canada supported by Atomic Energy Control Board, the Ontario Ministry of Labour, and the Workers' Compensation Board (WCB) in developing appropriate linkage methodology and data collection procedures to enable comprehensive follow-up studies of industrial cohorts, such as persons exposed to radiation (Fair *et al.* 1988a and 1988b). The co-operation of Employment and Immigration is also acknowledged.

Organizationally, the paper is divided into six parts:

1. A description is given of the background and opportunity that we had to conduct this study;

[1] M.E. Fair and P. Lalonde, Statistics Canada, R.H. Coats Building, 18th Floor, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

2. The main results and conclusion of this study are examined;

3. The files used are described;

4. The procedures used in the study are given;

5. The detailed results are presented; and

6. The future implications of the study in terms of:

   - estimating the anticipated error rates when two files are being linked;
   - using an 'intermediate' file to facilitate linkage;
   - developing recommended data collection procedures for studies, are discussed.

The typical kinds of questions that we are trying to address are:

1. What is the loss in accuracy if one only has birth year rather than full birth date (year, month, day) on the files being linked?

2. How much will the accuracy of linkage improve if one adds items such as birthplace and mother's maiden surname in the data collection?

3. What is the effect on the accuracy if there are only the initials and birth year rather than full forenames and birth date?

4. How important is it to retrieve complete forenames and birth dates rather than simply requesting initials and age on a survey if the records will later be linked to other data sources?

5. If the anticipated linkage accuracy is low, will linking to an 'intermediate' supplementary file such as the Social Insurance Number (SIN) index file, help to improve the quality of the match?

6. Does use of the Social Insurance Number master file identifiers substantially improve the accuracy of the death match? If so, how should this information be incorporated into the death search?

2. BACKGROUND AND THE OPPORTUNITY TO CONDUCT THIS STUDY

Society and legislators are often faced with urgent questions regarding the long-term health effects of possible harmful agents in the workplace and elsewhere. For example, the recent Chernobyl incident caused international interest by the media and the public regarding the potential health effects of radiation to the individuals exposed and possibly to their offspring. The health effects remote from the accident were also of world-wide concern.

Today there are strong pressures from society to determine and reveal the health risks to which the public is exposed, especially where the harm is cumulative or latent for an extended period of time, as with cancer. This pressure is coming from the media, regulatory agencies, organized labour, special commissions, and researchers. There is a real or perceived right of the worker, and members of the general public, to protection from known and as yet unestablished health hazards.

The kinds of statistics we are talking about here are quite different from those available from census or survey data, which are essentially snapshots at a point in time. In long-term medical follow-up, one must identify a study population and follow them for perhaps 20-30 years to find out their vital and health status. In order to have enough power to statistically detect such possible effects on health, it may be necessary to follow up large groups of individuals, to combine and compare results from several studies to see whether they are reproducible, and perhaps to pool data internationally.

Epidemiologists and statisticians doing historical mortality follow-up studies among industrial cohorts must almost be data detectives in order to get the kinds of data required for these sorts of studies. Vital statistics and health information systems which have been developed for administrative purposes and the resulting data bases constitute powerful sources of information for assessing health outcomes (Last 1986).

Over the past few years, computer-based record linkage has been increasingly used in Canada to conduct long-term medical follow-up studies. More than fifty investigations have been completed using the Canadian Mortality Data Base (Smith and Newcombe, 1980). This is an historic file containing over 6 million records for deaths occurring in Canada since 1950. This file contains the underlying coded cause of death. The Mortality Data Base, the Generalized Iterative Record Linkage system and the probabilistic matching techniques used have been described in detail earlier (Smith 1981; 1986; Hill and Pring-Mill 1985; Howe and Lindsay 1981).

The cornerstone for most historic cohort mortality studies is the assessment of a possible excess risk of death for the work force over an observation period. This evaluation requires combined information on the number of cohort members, the period of observation, and the expected numbers of deaths for all causes or specific causes. The cohort of workers is followed and the vital status determined at the end of the observation period (Figure 1) (Simons and Toulbee, pgs 25-90; Working Group on Health and Safety of Laboratory Workers; Redmond et al 1969; Monson 1980). The percentage of death certificates obtained for those known deaths, the proportions 'lost to follow-up' (that is the fraction of individuals identified at the outset but with unknown vital status at the end of the study) and the proportion of the cohort enumerated should be reported as part of the study results. These proportions bear directly on the reliability of the conclusions derived since they describe the amounts of missing data in the study. The proportion of the enumerated cohort that are 'lost to follow-up' is an upper bound on the number of unknown deaths.

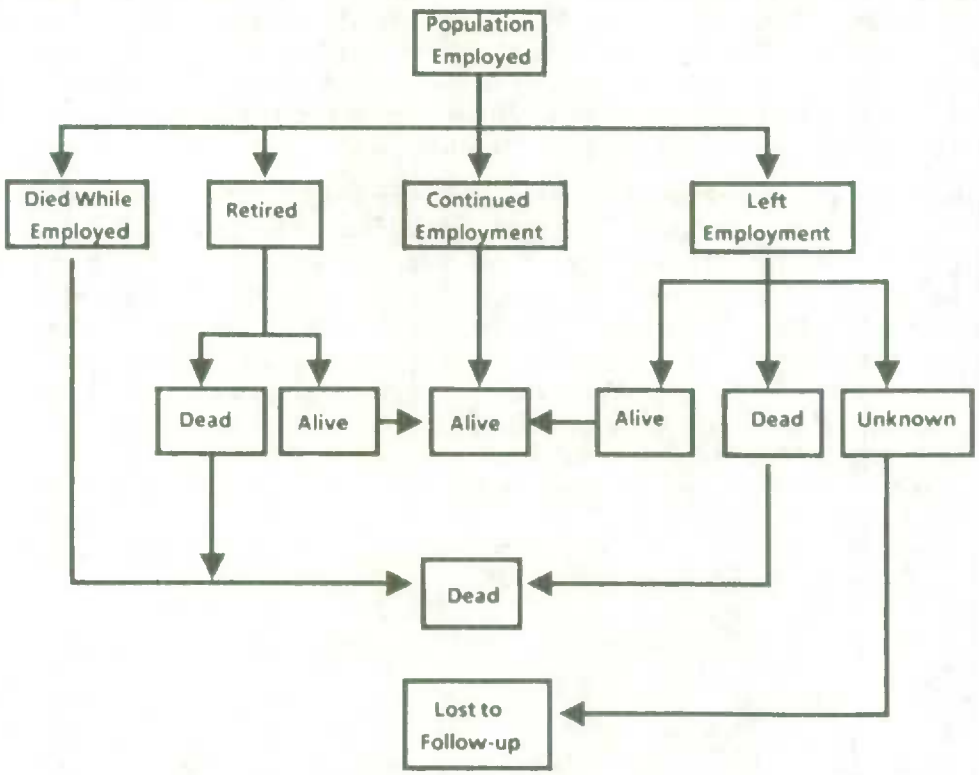


Figure 1. Follow-up of an industrial cohort to determine their vital status

Researchers doing studies are apt to vary widely in the "quality" of the linkage product they require. For some statistical purposes, a crude estimate of the proportions of deaths is sufficient. For other applications, it is frequently important that doubtful death linkages be resolved, and that any circumstances which preclude a death linkage, such as migration out of the country, be identified wherever possible. Epidemiologists doing cohort studies, particularly where the setting of regulatory standards and the development of compensation criteria are involved, often require intensive individual follow-up to determine the vital status of the cohort. It is desirable to trace as high a percentage of the cohort as possible. Failure to ascertain vital status for any appreciable segment of a study group may lead to erroneous or misleading conclusions. Questions may be raised about the possibility of bias in the results if the degree of follow-up is less than 95 per cent.

The tests for this present study were carried out using results from an earlier 1950-77 mortality search for a cohort of Ontario miners (Muller et al., 1983). The findings of our tests have had implications in the design and development of appropriate methodology for several other studies, particularly for industrial cohorts where the Social Insurance Number is available, and in the development of our current recommended data collection package which is designed for researchers wishing to carry out mortality, cancer or genetic studies in the future.

It is important to emphasize that no unique individual numeric identifier was present on the original two files being matched, and hence names, birth dates and such had to be used. The Social Insurance Number is rarely reported on Canadian death records in machine readable form, and in some provinces there is no place for it to be collected on the original source document. Better personal identifiers were believed to exist in the index files of the Social Insurance Number system. Use of this more reliable source of information as an 'intermediate' file for linkage was therefore considered likely to improve the death searches wherever the Social Insurance Number was available on the miners cohort file. The present study was undertaken to test the feasibility of the proposed use of the SIN master index identifiers, and to determine how much improvement in the accuracy of the death search would result. We used the Generalized Iterative Record Linkage system and probabilistic matching techniques for linking the Ontario miners data file to the Mortality Data Base.

In our tests, we were particularly interested in evaluating the quality of our statistical product in relation to the number of personal identifiers available on the file to carry out the linkage (Table 1). In the past we have recommended that full birth and current surnames, forenames, birth date and birth place be collected for files being matched against the Mortality Data Base. Mother's maiden surname and parental variables are also very useful. In particular, we wanted to obtain some data regarding the penalty for omission of particular identifiers. Birth dates and given names are known to be extremely important, but in the past we have had little quantitative data on the magnitude of the errors created when the various identifiers are not available on both files. We also wanted to investigate the use of the Social Insurance Number index master file to facilitate the death searches for the Ontario miners study, because if these variables were superior we could improve the quality of the output and reduce the amount of manual resolution required for the study.

Table 1

Identifiers on the WCB, SIN and MDB Records

Identifier	WCB Rec	SIN Rec	MDB Rec
Social Insurance Number	+	+	n.a.
Names*	+	+	+
Birth date	+	+	+*
Birth place	+**	n.a.	+*
Sex	+	+	+
Province of residence	+	n.a.	+
Mother's maiden surname	n.a.	+	+*
Year last known alive	+	n.a.	+

\+ Present on the file
* Not always complete
** Canada and the U.S.A. grouped together under the same code--not very discriminating
n.a. = not available.

3. MAIN RESULTS AND CONCLUSIONS

Of 30,000 miners' records which had previously been matched to the Canadian Mortality Data Base, 2,243 are judged to have been correctly matched with the corresponding death records. Of the 2,243 correctly matched records , 705 had the 'full' set of identifiers represented on both records in each of the pairs, i.e. full date of birth, two full given names, and mother's maiden surname. To investigate the extent to which the error rates are influenced by the availability or non-availability of the various identifiers on the records, we tried suppressing certain of the more important identifiers contained on the records singly and in groups. The deaths searches were then repeated, with these identifiers present and with certain of them removed or not used.

The procedures for carrying out the probabilistic death linkage results in a total weight. This is derived from a sum of the weights for agreement, partial agreement, or disagreement of the items compared. A threshold is set to determine at which point linkages will be accepted. For the present test, an 'optimum' threshold weight was defined as the level which ensures that the false positives and false negatives come closest to being equal in number. With this 'optimum' threshold, the total errors (i.e. the false positives plus false negatives) tend to be reasonably close to a minimum. The results of the death search are shown in Table 2. Clearly, the full date of birth is of special importance and full given names come next. Even a seemingly less important identifier such as middle name or initial, which is often not present on survey records, still makes a major difference to the accuracy of the death searches. Even the less frequently used identifiers such as mother's maiden surname, seem to become increasingly important where there is a shortage of the more frequently used identifiers (see lines 5-8 versus 1-4 in Table 2). Table 3 demonstrates the effects on the error rates when identifiers are deleted from the search record, and where the identifiers on the file could be either present or absent.

For those who may still wonder why even higher levels of accuracy of the death searches are not readily achieved, the answer lies in the quality and availability of the identifying variables contained in the original records. Some of the identifying information gets wrongly recorded. It should be emphasised that these data refer only to male miners.

Table 2

Effects on the Error Rates when Identifiers are Deleted from the "Hybrid" Search Records[1]

RUN #	Birth Y	Birth M	Birth D	Given Names I1	Given Names R1	Given Names I2	Given Names R2	MM	BP	Best Threshold	Counts False Pos	Counts False Neg	Counts Total Error	Total Errors as a % of 705 Good Link
Deletions of Birth Date Information														
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
2	+	+	-	+	+	+	+	+	.	46 +	12	12	24	3.4
3	+	-	-	+	+	+	+	+	.	50 +	23	23	46	6.5
4	-	-	-	+	+	+	+	+	.	37 +	45	45 + 3*	93	13.2
5	+	+	+	+	+	+	+	-	.	37 +	14	14	28	4.0
6	+	+	-	+	+	+	+	-	.	25 +	18	19	37	5.2
7	+	-	-	+	+	+	+	-	.	24 +	33	34 + 1	68	9.6
8	-	-	-	+	+	+	+	-	.	4 +	76	78 + 9*	163	23.1
Deletions of Name Information														
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
9	+	+	+	+	+	-	-	+	.	45 +	15	14	29	4.1
10	+	+	+	+	-	-	-	+	.	45 +	20	20	40	5.7
5	+	+	+	+	+	+	+	-	.	37 +	14	14	28	4.0
11	+	+	+	+	+	-	-	-	.	19 +	20	21 + 1*	42	6.0
12	+	+	+	+	-	-	-	-	.	7 +	25	25 + 1*	51	7.2
Deletions of Mother's Maiden Surname														
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
5	+	+	+	+	+	+	+	-	.	37 +	14	14	28	4.0

Table 2

(Continued) Effects on the Error Rates when Identifiers are Deleted from the "Hybrid" Search Records[1]

RUN #	Birth Y	Birth M	Birth D	Given Names I1	Given Names R1	Given Names I2	Given Names R2	MM	BP	Best Threshold	Counts False Pos	Counts False Neg	Counts Total Error	Total Errors as a % of 705 Good Link
	Deletions of Birthplace													
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
13	+	+	+	+	+	+	+	+	-	49 +	8	8	16	2.3
	Multiple Deletions of Identifiers													
1	+	+	+	+	+	+	+	+	.	50 +	8	8	16	2.3
3	+	-	-	+	+	+	+	+	.	50 +	23	23	46	6.5
15	+	-	-	+	+	-	-	+	.	43 +	41	41 + 1*	83	11.8
16	+	-	-	+	+	-	-	+	-	41 +	40	41 + 2*	83	11.8
17	+	-	-	+	+	-	-	-	.	1 +	70	66 + 5*	141	20.0
20	+	-	-	+	-	-	-	-	.	- 14 +	98	96 + 8*	202	28.6
18	-	-	-	+	+	-	-	+	.	21 +	87	89 + 10*	186	26.4
19	-	-	-	+	+	-	-	-	.	- 19 +	144	143 + 32*	319	45.2
14	+	+	+	+	+	+	+	-	-	36 +	13	13 + 1*	27	3.8

* See Footnotes for Tables 2 and 3

[1] Based on just those search records from the correctly matched pairs having a full set of identifiers on both records -- i.e. full date of birth, two full given names, and mother's maiden surname. Birthplace was present or absent on both records. Percentages are calculated on 705 good links.

Key to Headings:

I = Initials of first (I1) or second (I2) given name
R = Remainder of first (R1) or second (R2) given name
MM = Mother's maiden surname
BP = Birth place
AMJ = Year, Month, Day

Key to Entries:

Identifier: Deleted (-); Present (+); Present or absent (.)

Table 3

Effects on the Error Rates when Identifiers are Deleted from the "Hybrid" Search Records[1]

RUN #	Birth Y	Birth M	Birth D	Given Names I1	Given Names R1	Given Names I2	Given Names R2	MM	BP	Best Threshold	Counts False Pos	Counts False Neg	Counts Total Error	Total Errors as a % of 2243 Good Links
	Deletions of Birth Date Information													
1	.	.	.	.	.	.	.	.	.	50 +	54	55	109	4.9
2	.	.	-	.	.	.	.	.	.	40 +	82	81 + 1*	164	7.3
3	.	-	-	.	.	.	.	.	.	38 +	125	123 + 10*	258	11.5
4	-	-	-	.	.	.	.	.	.	16 +	229	233 + 41*	503	22.4
	Deletions of Name Information													
1	.	.	.	.	.	.	.	.	.	50 +	54	55	109	4.9
5	.	.	.	.	.	-	-	.	.	42 +	61	63 + 1*	125	5.6
6	.	.	.	.	-	-	-	.	.	35 +	79	81 + 1*	161	7.2
	Deletions of Other Identifiers													
7	.	.	.	.	.	.	.	.	-	40 +	45	46	91	4.1
8	.	.	.	.	.	.	.	-	.	35 +	52	51	103	4.6

* See Footnotes for Tables 2 and 3

[1] Based on using all 30,000 potentially linkable search records. Percentages are calculated on 2,243 good links.

Key to Headings:

I = Initials of first (I1) or second (I2) given name
R = Remainder of first (R1) or second (R2) given name
MM = Mother's maiden surname
BP = Birth place

Key to Entries:

Identifier: Deleted (-); Present (+); Present or absent (.)

Footnotes Regarding Tables 2 and 3

1.* In Tables 2 and 3 the "false negatives" with an asterisk are due to potentially linkable search records becoming matched preferentially with the wrong death records; the weights for these will normally be below the "best" threshold so that they are unlikely to become "false negatives" that are due solely to the weight for a correct match falling below the "best" threshold.

2. The "best" threshold is taken to be the threshold at which the "false positives" and the "false negatives" are most nearly equal in number.

3. Of 30,000 miners' records, 2243 are judged to have been correctly matched with the corresponding death records as a result of the "hybrid" search. Of the 2243 correctly matched pairs, 705 had the "full" set of identifiers represented on both records in each of the pairs. Table 2 is based on the death records from the 705 pairs with full identifiers, and Table 3 is based on the 2,243 correctly matched pairs from the file of 30,000 potentially linkable search records.

4. THE FILES USED

The study cohort of Ontario miners is represented by about 50,279 records from the Workers' Compensation Board. Of these, a total of precisely 30,000 contained valid Social Insurance Numbers and were therefore linkable to the Social Insurance Number index file. Records from both these sources are potentially linkable with the death records contained in Canada's Mortality Data Base using names and such.

The vital status of the 30,000 miners whose SIN numbers are known had been ascertained earlier by a particularly thorough process consisting of: a) a death search, b) an 'alive' search of 1977-78 tax records, and c) a manual check of any apparent conflicts to uncover the reason and the true status.

From the prior 1950-77 death searches, at an intermediate COMPARE phase in the linkage operation utilizing the Generalized Iterative Record Linkage system, a comparatively small file of death records was saved which included all that were even remotely likely to link with the miners records either correctly or falsely (see prod7-DATB file in Table 4). Since the Social Insurance Number was introduced in 1964, the file was used starting with deaths from 1964 onward. The records had to agree on the New York State Identification and Intelligence System phonetic code known as 'NYSIIS' to be included in this truncated death file.

Table 4

Number of Miners' WCB Records and Number of Death Records Used for the "SIN Evaluation" Study

Records	Totals
Miners' WCB Records (males)	
Total in original file	50,279
Total with valid SIN Numbers	30,000
Vital status of the 30,000 miners with SIN numbers:	
Confirmed alive	26,736
Death links (1964-77) confirmed	2,254
Lost to follow-up	1,010
Death Records Searched (males)	
Total in MDB (1964-1977) approx.	1,300,000
Total in "prod7-DATB" file (1950-1977)	46,679
Total in "prod7-DATB" file (1964-1977)	35,251

5. PROCEDURES OF THE STUDY

For the present study use was made of four files, a death file plus three sets of records pertaining to Ontario miners. The latter were used to initiate the death searches. More specifically, these files consisted of:

1. Death records from the 1964-77 CMDB that had survived the COMPARE phase of an earlier death search;
2. WCB records with SIN numbers;
3. SIN index records with the same SIN numbers as the WCB records; and
4. Composite WCB-SIN hybrid records.

Of the 30,000 WCB records with SIN numbers, only 2,254 related to miners who were actually known to be dead. However, all 30,000 have been included in the test since some of them could conceivably contribute to the numbers of false linkages. For each WCB record used in the test, the corresponding record with the same SIN index number was drawn from the SIN file. These records were used, both separately and together.

The plan for the present study was to repeat the death searches using:

1. the WCB identifiers alone,
2. the SIN identifiers alone, and
3. a composite or 'hybrid' record containing identifiers from both.

The 'hybrid' record was designed to contain the mother's maiden surname from the SIN record because it is not recorded in the WCB record, and the birth place code from the WCB record because it is absent from the SIN record. Where the personal identification (i.e. surname, year, month, day of birth, and two given names) was represented differently on the two source records, then two separate entry records were created. If all six items were identical, then only one record was kept. The personal identification was different in nearly 60% of the cases.

6. RESULTS — ACCURACIES OF THE DEATH SEARCH

The study was designed primarily to answer a simple question i.e. "Does use of full personal identifiers such as those on the SIN index substantially improve the accuracy of the death link?" The simple answer to this question is that the use of the SIN identifiers greatly increases the capture of the known potential correct death links (see Tables 5 and 6). With the WCB identifiers alone, 56 of the 2254 potential good links were missed entirely; with the use of the SIN index file plus WCB identifiers the number of such losses was reduced to just 7 (i.e. 0.3 per cent as against 2.5).

Threshold weights have to be selected in completing a linkage. As mentioned earlier, we selected an 'optimal' threshold i.e. a level where the false positives and false negatives came closest to being equal in number. To find the 'optimum' threshold, the matched pairs of records (i.e. the best death match for each miners record) were arrayed in descending order of total weight. We used a breakdown by ranges of ten units of weight and a finer breakdown of these weights in the vicinity of zero. From these distributions, a precise 'optimum' threshold may be derived for the death searches initiated in each of the three ways (Table 7). Such 'optimum' thresholds may vary with the amount of identifying information available. They also tend to fall somewhat above the theoretical 50-50 odd point (total weight = zero).

Using the 'optimum' weight, one may determine the numbers of false linkages that are accepted because their total weights are above that threshold, and the numbers of potential good matches that are rejected because their total weights are below the threshold.

The tests to determine the effects on the accuracy of searching, when various personal identifiers are deleted, serve to emphasize the value of using a source of these identifiers in which they are of high quality and completeness.

Table 5

Number of "Best" Death Matches by Type of Search Record — Dead Only

(The searches all relate to the 2,254 miners known to have died in the period 1964-77.)

Kind of Match Achieved	Number of Search Records		
	WCB Search Record	SIN Search Record	Hybrid Search Record
Counts			
Correct match was made; it was the "best"	2,193	2,238	2,243
Correct match made; it was not the "best"	5	3	4
No match was made	56	13	7
Total search records	2,254	2,254	2,254
Percentages			
Correct match was made; it was the "best"	97.29	99.29	99.51
Correct match made; it was not the "best"	.22	.13	.18
No match was made	2.48	.58	.31
Total search records	100.00	100.00	100.00

NOTE: The "best" match is by definition the one with the highest calculated odds in favour of a correct death linkage.

Table 6

Number of Correct Links Found, by Type of Search Record (Based on the same data as Table 5)

Correct Links Found by			Number of Correct Links Found*	% of Correct Links Found*
WCB Search Record	SIN Search Record	Hybrid Search Record		
Yes	Yes	Yes	2,192	97.25
No	Yes	Yes	49	2.17
Yes	No	Yes	6	.27
No	No	No	7**	.31
Total			2,254	100.00
Combined				
Yes	--	--	2,198	97.52
--	Yes	--	2,241	99.42
--	--	Yes	2,247	99.69

* All correct links are included, including the few that carried lower weights than an alternative incorrect death match.

** The 7 potential correct death links that were not detected by any of the three kinds of search records included:

4 due to differences in the surname NYSIIS code,
2 due to late registration of the death,
1 due to the death being outside the country (i.e. Germany).

Table 7

False Positive Links and False Negative Outcomes, Using an "Optimum" Threshold

Search Records (Number potentially Linkable)		Optimum Threshold	False Pos		False Neg		Total	
			No.	(%)	No.	(%)	No.	(%)
Based on Search Records that Formed Matched Pairs								
(Excludes the known truly linkable search records that did not form matched pairs in this operation.)								
WCB records	(2198)	+15	80	(3.6)	83	(3.8)	163	(7.4)
SIN records	(2241)	+16	55	(2.5)	54	(2.4)	109	(4.9)
Hybrid rec.	(2247)	+50	55	(2.4)	54	(2.4)	109	(4.9)
Based on all Search Records								
(Includes the known truly linkable search records that did not form matched pairs in this operation.)								
WCB records	(2254)	+15	80	(3.5)	139	(6.2)	219	(9.7)
SIN records	(2254)	+16	55	(2.4)	67	(3.0)	122	(5.4)
Hybrid rec.	(2254)	+50	55	(2.4)	61	(2.7)	116	(5.1)

7. FUTURE IMPLICATIONS OF THE STUDY

The results of this study are applicable to future studies of the mortality experience of occupational cohorts, including many studies relating to the long-term health effects of radiation. The improvement in the linkage will be marked where full identifiers are collected.

The use of 'intermediate' files to facilitate the linkage has proved to be particularly useful. It has been convenient to create from these a composite or hybrid record containing fields from both sources to complete the search. Other kinds of 'intermediate' records that could be used would for example include marriage records to facilitate the linkage of women for studies where name changes may have occurred during the period of follow-up.

This information is often required by persons:

1. Carrying out forms revisions (e.g. in the current review of vital statistics records);
2. Recommending and/or carrying out data collection (e.g. for occupational and environmental health research studies);
3. Designing questionnaires (e.g. for surveys, cancer registries);
4. Evaluating potential biases in the analysis of study results (e.g. how to handle 'lost to follow-up' cases in mortality studies);
5. Estimating the feasibility and possible error rates for a new study, given the availability of identifying items; and
6. Considering the improvement in accuracy that could result from using an 'intermediate' file to facilitate the match.

ACKNOWLEDGEMENT

The authors thank Dr. Newcombe for his assistance with this paper.

REFERENCES

Fair, M.E., Newcombe, H.B., Lalonde, P., and Poliquin, C. (1988). "Alive" searches as complementing death searches in the epidemiological follow-up of Ontario miners. A research paper prepared by Statistics Canada under contract to Atomic Energy Control Board. (In preparation for publication, Atomic Energy Control Board 1988a)

Fair, M.E., Newcombe, H.B., and Lalonde, P. (1988). Improved mortality searches for Ontario miners using Social Insurance Index identifiers. A research paper prepared by Statistics Canada under contract to Atomic Energy Control Board. (In preparation for publication, Atomic Energy Control Board 1988b).

Hill, T., and Pring-Mill, F. (1985). Generalized iterative record linkage system. In *Record Linkage Techniques - 1985. Proceedings of the Workshop on Exact Matching Methodologies*, (Eds., B. Kilss and W. Alvey), Dept. of the Treasury, Internal Revenue Service, 327-333.

Howe, G.R., and Lindsay, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.

Last, J.M. (1986). Individual privacy and health information: an ethical dilemma? *Canadian Journal of Public Health*, 77, 168-169.

Monson, R.A. (1980). *Occupational Epidemiology*. Boca Raton, Florida: CRC Press, Inc.

Muller, J., Wheeler, W.C., Gentleman, J.F., Suranyi, G., and Kusiak, R.A. (1983). Study of Mortality of Ontario Miners, 1955-1977 Part I. Ontario Ministry of Labour, Special Studies and Health Services Branch, 400 University Avenue, Toronto, Ontario, M7A 1T7.

Redmond, C.K., Smith, E.M., Lloyd, J.W., and Rush, H.W. (1969). Long-term mortality study of steelworkers. III Follow-up. *Journal of Occupational Medicine*, 11: 513-521.

Smith, M.E., and Newcombe, H.B. (1980). Automated follow-up facilities in Canada for monitoring delayed health effects. *American Journal of Public Health*, 70, 1261-1268.

Smith, M.E., and Silins, J. (1981).Generalized iterative record linkage system. *Social Statistics Section, Proceedings of the American Statistical Association*, 128-137.

Smith, M.E. (1981). Long-term medical follow-up in Canada. *Quantification of Occupational Cancer. Banbury Report No. 9.* Cold Spring Harbor, New York, Cold Spring Harbor Laboratory, 675-688.

Smith, M.E. (1986). Future needs and directions for computerized record linkage in health research in Canada. a) Future study plans and studies using the Canadian mortality data base. In *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, (Eds., G.R. Howe and R.A. Spasoff) U. of Toronto Press, 211-230 and 249-257. Proceedings are available from: Dept. of Epidemiology and Community Medicine, U. of Ottawa, Health Sciences Building, Ottawa, Ontario K1H 8M5.

Symons, M.J., and Taulbee, M.J. (1984). Statistical evaluation of the risk of cancer mortality among industrial populations. In *Statistical Methods for Cancer Studies*, 270 Madison Avenue, New York, Marcel Dekker Inc., 25-90.

Working Group on Health and Safety of Laboratory Workers (1986). Health and safety of lab workers in Canada - A code of Conduct. Occupational Health in Ontario 7,2, 92-108.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

COMPUTATIONAL ASPECTS OF APPLYING OF THE FELLEGI-SUNTER MODEL OF RECORD LINKAGE TO LISTS OF BUSINESSES

WILLIAM E. WINKLER[1]

ABSTRACT

Let **AXB** be the product space of two sets **A** and **B** which is divided into **matches** (pairs representing the same entity) and **nonmatches** (pairs representing different entities). Linkage rules are those that divide **AXB** into **links** (designated matches), **possible links** (pairs for which we delay a decision), and **nonlinks** (designated nonmatches). Under fixed bounds on the error rates, Fellegi and Sunter (1969) provided a linkage rule that is optimal in the sense that it minimizes the set of possible links. The optimality is dependent on knowledge of certain joint inclusion probabilities that are used in a crucial likelihood ratio. In applying the record linkage model, assumptions are often made that allow estimation of the joint inclusion probabilities. If the assumptions are not met, then a record linkage procedure using estimates computed under the assumptions may not be optimal. This paper contains an examination of methods for adjusting linkage rules when assumptions are not valid. The presentation takes the form of an empirical analysis of lists of businesses for which the truth of matches is known. The number of possible links obtained using standard and adjusted computational procedures may be dependent on different samples. Bootstrap methods (Efron, 1987) are used to examine the variation due to different samples.

1. INTRODUCTION

This paper presents an analysis of the computational aspects of applying the Fellegi-Sunter model of record linkage to lists of businesses.

Given two lists, we wish to use identifying information to delineate those record pairs that represent the same entities (**matches**) and those that are different (**nonmatches**). Thus, we desire to define a linkage rule that allows us to divide the crossproduct space of pairs into **links** (designated matches), **possible links** (pairs for which a decision is delayed), and **nonlinks** (designated nonmatches).

In implementing the model, the basic idea is to use computer procedures to designate automatically links and nonlinks. Manual review is then used to gather (sometimes) more information about possible links and designate them as links or nonlinks.

[1] William E. Winkler, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233, USA.

Under fixed bounds on the numbers of erroneous matches and nonmatches, Fellegi and Sunter (1969, Theorem) provide a procedure that, in theory, minimizes the number of possible links. The optimality is dependent on knowledge of certain joint inclusion probabilities that are used in a crucial likelihood ratio.

The Fellegi-Sunter model is particularly useful because it allows progressive improvements in how software delineates and compares information to be incorporated in record linkage procedures. For instance, assume the following pair is designated as a

Name	Street	City	State	ZIP
Zabrinsky, Robert A	16 Sycamore St	Dayton	OH	53342
Zabrinsky, R	167 W Sycamore	Dayton	OH	53342

possible link. The pair has been obtained by a procedure that uses an abbreviation of the surname. Existing software only allows comparison of the remaining portions of the address on a character by character basis.

If software is developed that allows comparison of 'SYCAMORE' with 'SYCAMORE', then the computerized procedure which incorporates the additional information might designate the pair as a link. In essense, we try to develop software that will parallel decisions a human might make.

Newcombe et al. (1983) provide an example of how computer procedures can outperform manual procedures. Their example involves population files having substantial amounts of information that can be efficiently compared.

Many lists, however, do not have information in a form that allows efficient comparisons. Thus, we wish to examine general methods that allow us to make progressively better decisions. Given fixed bounds on error rates, **better** linkage rules will be those that reduce the set of possible links. Such linkage rules can be obtained when additional software allows use of more of the available information or adjusts computational procedures when assumptions are not entirely valid.

The remainder of the paper presents methods for applying the Fellegi-Sunter model to lists of businesses. The presentation takes the form of an analysis of sizes of the regions of possible links under fixed bounds on errors rates. The application involves pairs of lists for which the truth and falsehood of linkages are known.

Although the computational methods used are specific to the types of lists encountered at the Energy Information Administration (EIA), generic similarity to methods that are used in other record linkage applications will be pointed out.

The second section of this paper is divided into four subsections. The first contains a description of the data base and the specific subfields that are compared. The second subsection contains both a summary of the basic Fellegi-Sunter model and the aspects that allow its ready modification in specific applications. The third subsection highlights common assumptions made and computational procedures used in record linkage projects in Canada and the United States. It also contains details of computational procedures that are specific to the application of this paper.

The fourth subsection describes the evaluation procedures. The basic evaluation technique involves comparing sizes of the region of possible links when different types of linkage rules are applied under fixed error bounds. The sizes of the regions of possible links are statistics that may be dependent on the samples used in calibrating the linkage rules. Efron's bootstrap (1987, 1982, 1979) is used to evaluate their distributions.

Results are presented in the third section. This is followed in the fourth section by discussion of additional types of comparisons and of use of additional blocking. Finally, the paper concludes with a summary.

2. DATA BASE, LINKAGE MODEL, COMPUTATIONAL AND EVALUATION PROCEDURES

This section contains descriptions of the data base, the Fellegi-Sunter model of record linkage, computational procedures, and evaluation methods.

2.1 Data Base

The description of the data base is divided into two components. The first component is a description of the overall properties. The second contains a listing of the specific subfield comparisons that are made.

2.1.1 Overall Description

The data base of 57,900 records contains 54,850 records that are identified as individual companies and 3,050 duplicates. A pair of records that consists of a company and its corresponding duplicate is a match; all others are nonmatches.

The data base was constructed from 11 EIA and 47 State and industry lists containing 176,000 records. Duplicates were identified via elementary techniques, through call-backs (phone numbers are sometimes present) and through surveying.

The purpose of this paper is to examine how the basic Fellegi-Sunter model can be modified to allow accurate delineation of hard-to-identify duplicates. Easily identified duplicates (those generally having substantial portions of their name and addresses agreeing on a character by character basis) are not considered.

An example of a hard-to-identify duplicate might be:

Name	Street	City	State	ZIP
Zabrinsky Fuel	16 W Sycamore St	Dayton	OH	53315
Zabrinky Cmpny	167 Sycamere St	Springfield	OH	53315

We observe that both 'Zabrinsky' and 'Sycamore' are spelled wrong in the second record, that 'Cmpny' is a nonstandard abbreviation, and that Springfield OH, a suburb of Dayton, has Postal ZIP code 53315.

2.1.2 Specific Subfields Compared

There are four sets of specific subfields that are compared in each pair of records. First are those that can be obtained through easy substring comparisons. The field designated WL-NAME given below is obtained by sorting the NAME field by words of decreasing length with ties broken by an alpha sort.

FIELD NAME	SUBFIELD COLUMNS COMPARED
NAME	1-4, 5-10, 11-20, 21-30
STREET	1-6, 7-15, 16-30
ZIP	1-3, 4-5
CITY	1-5, 6-10, 11-15
STATE	1-2
TELEPHONE	1-3, 4-6, 7-10
WL-NAME	1-4, 5-10, 11-20, 21-30

The second set is the four comparisons of the first and second largest words in the NAME field. Ties are again broken by an alpha sort.

The last two sets are of subsets of the STREET and NAME fields that are designated by highly sophisticated software. ZIPSTAN software from the Census Bureau (U.S. Dept. of Commerce 1978b) is used to obtain corresponding subfields of the STREET field. The subfields are: House No., Prefixes 1 and 2, Street Name, Suffixes 1 and 2, and Unit. Prefixes are directions such as East and North. Suffixes are words such as Street and Road. Unit designates identifiers such as apartment or suite number.

The NSKGEN5 module from software used in the Canadian Business Register (Statistics Canada 1984, 1982) is used to obtain corresponding subfields of the NAME field. NSKGEN5 creates three groups of words. The first group consists of three abbreviations with the first corresponding to SURNAME if present. The second group contains two words with the first corresponding to surname. The third group is a single word obtained by concatenating and abbreviating individual words in the NAME field. Details are given in Winkler (1987a) or in Statistics Canada (1984, 1982).

Intuitively, a given comparison of corresponding subfields of the NAME field may be dependent on comparison of different corresponding subfields of the NAME field.

2.2 Fellegi-Sunter Model

The Fellegi-Sunter Model uses a decision theoretic approach establishing the validity of principles first used in practice by Newcombe (Newcombe et al. 1959). To give an overview, we describe the model in terms of ordered pairs in a product space. The description closely follows Fellegi and Sunter (1969, pp. 1184-1187).

There are two populations **A** and **B** whose elements will be denoted by a and b. We assume that some elements are common to **A** and **B**.

Consequently the set of ordered pairs

$$AXB = \{(a,b): a\varepsilon A,\ b\varepsilon B\}$$

is the union of two disjoint sets of **matches**

$$M = \{(a,b): a=b,\ a\varepsilon A,\ b\varepsilon B\}.$$

and **nonmatches**

$$U = \{(a,b): a\neq b,\ a\varepsilon A,\ b\varepsilon B\}.$$

The records corresponding to **A** and **B** are denoted by $\alpha(a)$ and $\beta(b)$, respectively. The **comparison vector** τ associated with the records is defined by:

$$\tau[(\alpha(a),\beta(b))] \equiv \{\tau^1[(\alpha(a),\beta(b))],\tau^2[(\alpha(a),\beta(b))],\ldots,\tau^K[(\alpha(a),\beta(b))]\}.$$

Where confusion does not arise, the function τ on **AXB** will be denoted by $\tau(\alpha,\beta)$, $\tau(a,b)$, or τ. The set of all possible realizations of τ is denoted by Γ.

The conditional probability of $\tau(a,b)$ if $(a,b)\varepsilon M$ is given by

$$m(\tau) \equiv P\{\tau[(\alpha(a),\beta(b))] \mid (a,b)\varepsilon M\}$$

$$\sum_{(a,b)\varepsilon M} P\{\tau[(\alpha(a),\beta(b))]\}.P[(a,b)\mid M].$$

Similarly we denote the conditional probability of τ if $(a,b)\varepsilon U$ by $u(\tau)$.

We observe a vector of information $\tau(a,b)$ associated with pair (a,b) and wish to designate a pair as a link (in set A_1), a possible link (in set A_2), or a nonlink (in set A_3). We let L denote a linkage rule that divides **AXB** into A_1, A_2, and A_3. We say that a **Type I error** has occurred if rule L places $m\varepsilon M$ in A_3 and a **Type II error** if L places $u\varepsilon U$ in A_1. Fellegi and Sunter (1969) define a linkage rule L_0, with associated sets A_1, A_2, and A_3 that is optimal in the following sense:

THEOREM (Fellegi-Sunter 1969). Let L' be a linkage rule with associated sets A_1', A_2', and A_3' such that $P(A_3'|M) = P(A_3|M)$ and $P(A_1'|U) = P(A_1|U)$. Then $P(A_2|U) \leq P(A_2'|U)$ and $P(A_2|M) \leq P(A_2'|M)$.

In other words, if L' is any competitor of L_0 having the same Type I and Type II error rates (which are both conditional probabilities), then the conditional probabilities (either on set U or M) of not making a decision under rule L' are always greater than under L_0.

The Fellegi-Sunter linkage rule is actually optimal with respect to any set Q of ordered pairs in **AXB** if we define error probabilities P_Q and a linkage rule L_Q conditional on Q. Thus, it may be possible to define subsets of **AXB** on which we make use of differing amounts and types of available information.

For instance, if we have a set of pairs in which telephone number is present, we might use telephone number and a few characters from the name to designate links. With other pairs, we may additionally have to utilize information from the street address and the city name.

Sets of ordered pairs Q on which the Fellegi-Sunter linkage rule is applied are often obtained by **blocking criteria.** Blocking criteria are sort keys that are used to reduce the number of pairs that are considered. Rather than consider all pairs in **AXB,** we use might only consider pairs that agree on the first three digits of the ZIP code or on a suitable abbreviation of surname.

2.3 Computational Procedures

The primary goals of record linkage are finding implementable computational procedures that effectively utilize information available in files, that allow reasonably straightforward updating with enhanced procedures, and that can be tested.

The computational procedures are divided into six parts. The first part contains a description of the general linkage rule of the Fellegi-Sunter Model. The second contains a description of the simplified computational procedures when a conditional independence assumption is made.

The computational procedures in parts one and two have been implemented (with minor variations) at the U.S. Bureau of the Census for population files (Kelley 1985a, 1985b, and 1986; U.S. Dept. of Commerce, 1978a), the U.S. National Death Index (Rogot et al. 1983), the U.S. Department of Agriculture (Coulter 1977; U.S. Dept. of Agriculture 1979), the California Automated Mortality Linkage System (Arellano 1985), the U.S. Energy Information Administration (Winkler 1985b, 1987a), and Statistics Canada (Smith and Silins 1981; Smith, Newcombe, and Dewar 1983).

Background on the validity of the conditional independence assumption is presented in the third part. The fourth contains an observation about equivalent linkage rules. The fifth describes two general methods of adapting computational procedures. The sixth provides a description of the specific computational procedures of this paper.

2.3.1 General Form of Linkage Rule

To provide a background for understanding why specific computational procedures are used, we consider the following likelihood ratio

$$R \equiv R[\tau(a,b)] = m(\tau)/u(\tau). \tag{2.1}$$

If the numerator is positive and the denominator is zero in (2.1), we assign an arbitrary very large number to the ratio. The Fellegi-Sunter linkage rule takes the form:

If $R > \text{UPPER}$, then denote (a,b) as a link.

If $\text{LOWER} \leq R \leq \text{UPPER}$, then denote (a,b) as a possible link. (2.2)

If $R < \text{LOWER}$, then denote (a,b) as a nonlink.

The cutoffs LOWER and UPPER are determined by the desired error rate bounds.

2.3.2 Simplification Under Conditional Independence Assumption

In practice, computation is simplified two ways. The first is by the conditional independence assumption of Fellegi and Sunter (1969):

$$m(\tau) = m_1(\tau^1).m_2(\tau^2)\ldots m_k(\tau^K) \text{ and}$$

$$u(\tau) = u_1(\tau^1).u_2(\tau^2)\ldots u_k(\tau^K)$$

where for $i = 1, 2, \ldots, K$

$$m_i(\tau^i) = P(\tau^i \mid (a,b) \varepsilon M) \text{ and}$$

$$u_i(\tau^i) = P(\tau^i \mid (a,b) \varepsilon U)$$

The second is to use a computationally convenient function of the ratio in (2.1). Log_2 is used. We then have

$$W \equiv W(\tau) = \text{Log}_2[m(\tau)/u(\tau)] \tag{2.3}$$

$$= W^1 + W^2 + \ldots + W^K,$$

where $W^1 \equiv \text{Log}_2\,[m_i(\tau^i)/u_i(\tau^i)]$ for $i = 1, 2, \ldots, K$. We call W the **total comparison weight** associated with a pair and W_i, $i = 1, 2, \ldots, K$, the **individual comparison weights.**

For the remainder of the paper we will assume that each component τ^i, $i = 1, 2, \ldots K$, in τ represents a two-state comparison (e.g., agree/disagree) and define the marginal comparison events by

$$B^i \equiv \{(a,b) \mid \tau^i(a,b) = \tau^i_0\}$$

for one fixed state of τ^i_0. Under the Conditional Independence Assumption we need to estimate 2K probabilities of the form

$$P(\tau \varepsilon B_i \mid M) \text{ and } P(\tau \varepsilon B_i \mid U), \; i = 1, 2, \ldots, K. \tag{2.4}$$

If we have a set of pairs for which truth and falsehood of matches are known, then, for each agreement characteristic B_i, $i = 1, 2, \ldots, K$, we need to divide it into the four subsets determined by (2.4) to perform the estimation.

If no assumptions are made, we need to estimate $2 \cdot 2^k - 1$ joint inclusion probabilities (the numerator and denominator in formula (2.1)) and divide the set of pairs for which truth and falsehood are known into $2 \cdot 2^k - 1$ subsets. Even with a small number of comparisons (say, 6 or less), we may not be able to obtain sufficiently large samples to allow accurate estimation of the joint inclusion probabilities.

In previous applications, Newcombe and Kennedy (1962) made more than 60 comparisons, Rogot et al. (1983) made 11, and Winkler (1985b, 1987a) made more than 30.

Fellegi and Sunter (1969) gave two methods that allow computation of probabilities of the form (2.4). Method I makes several assumptions that are based on prior knowledge of file characteristics. Method II allows computation of probabilities of the form (2.4) directly from file characteristics and does not require knowledge of truth of matches obtained from samples.

2.3.3 Validity of Conditional Independence Assumption

Fellegi and Sunter indicate that, if the Conditional Independence Assumption is not valid, then estimates of weights that are obtained via formula (2.3) will lose their strict probabilistic interpretation. By this, they mean that the linkage rule of their theorem may not actually minimize the number of possible links. They indicate that they believe their procedure to be robust to departures from the independence assumption.

Winkler (1985b) has shown that the independence assumption is not valid for simple comparisons of portions of the name and street address fields for list of businesses. Using similar portions of the name and street fields, Kelley (1986) has shown that the independence assumption is not valid for files of individuals. Furthermore, Kelley and Winkler have each shown that matching efficacy is sensitive to the set of pairs over which probabilities of the form (2.4) are computed.

Under the independence assumption, joint inclusion probabilities are computed as products of probabilities of the form (2.4). If we have a set of pairs for which truth and falsehood are known, then we can adjust the joint inclusion probabilities for departures from the independence assumption. If the adjustments are substantial, then the Fellegi-Sunter procedure may not be robust to departures from independence.

2.3.4 Equivalent Rules

We observe that, for each set of comparisons $B_1, B_2, \ldots, B_K$ (and associated comparison space $\Gamma = \{\tau(a,b): \; (a,b) \; \varepsilon \; A \times B\}$), the set of weights of the form (2.3) induce a linear ordering in Γ. Ties are ordered arbitrarily. Any other set of real numbers that are assigned to Γ and that yield the same linear ordering will yield optimal linkage

rules of the form (2.2) **for every pair of error bounds** for which the original Fellegi-Sunter linkage rule is valid.

If the pair of error bounds are fixed, then we only need to assign real numbers R', cutoffs LOWER' and UPPER', and linkage rules of the form (2.2) so that the sets of links and nonlinks agree with those given by the optimal Fellegi-Sunter rule. Consequently, we have considerable latitude in how we actually implement linkage rules.

2.3.5 General Adjustments

There are two general adjustments to the basic methods of computing individual comparison weights. The first consists of dividing the subset of pairs in **AXB** over which individual comparison weights are computed into several subsets. The linkage rule is obtained by restricting the basic Fellegi-Sunter rule to correspond to the different subsets on which weights are computed. Individual comparison weights may vary significantly in different subsets.

The second adjustment consists of modifying individual comparison weights. If, under the independence assumption, we consider the equation

$$W \equiv \mathrm{Log}_2(P(\tau\varepsilon B_i \cap B_2 \cap \ldots \cap B_K | M)/P(\tau\varepsilon B_1 \cap B_2 \cap \ldots \cap B_k | U))$$

$$= W^1 + W^2 + \ldots + W^K,$$

where $W^i \equiv \mathrm{Log}_2\ (P(\tau\varepsilon B_i | M)/P(\tau\varepsilon B_i | U))$ for $i = 1, 2,$ and K, then we wish to find computationally tractable methods of adjusting the W^i, $i = 1, 2, \ldots, K$, so that their sum yields better linkage rules.

If there is a sample for which the truth and falsehood of matches are known, then we can estimate individual comparison weights (Tepping 1968) and the adjustments.

The simplest adjustment procedure involves a steepest ascent approach (e.g., Cochran and Cox 1957). To begin, we use the known truth and falsehood of matches within a sample to estimate probabilities of the form (2.4). The probabilities are then used in computing individual comparison weights that are added to obtain an estimate of total weight (2.3). For each pair of fixed bounds on Type I and Type II errors, the cutoffs UPPER and LOWER of (2.2) can be determined. The number of potential links for rules of the form (2.2) follows immediately.

Next, we chose an individual comparison weight, change it by a fixed amount (say $\pm$ 1), recompute the total weight of (2.2) using the new individual weight, and find new cutoffs UPPER and LOWER and a new region of potential links.

If under fixed bounds of errors, the size of the region of possible links decreases, then we continue adjusting the individual comparison weight (either up or down) until the region ceases its decrease in size. We continue by varying other individual weights in a similar manner.

If the size of the region of possible links decreases substantially, then we know the Conditional Independence Assumption is not valid for the set of comparisons.

A linkage rule that is based on adjusted individual comparison weights depends on the sample used in the steepest ascent procedure.

2.3.6 Specific Methods

To describe the specific methods of computing weights and obtaining corresponding linkage rules, we need some additional background.

The only sets of pairs considered are those that agree on the following blocking criteria.

Blocking Criteria
1. 3 digits ZIP, 4 characters NAME
2. 5 digits ZIP, 6 characters STREET
3. 10 digits TELEPHONE
4. Word length sort NAME field, then use 1*

* This criterion also has a deletion stage which prevents matching on commonly occurring words such as 'OIL,' 'FUEL,' 'CORP,' and 'DISTRIBUTOR.'

We subdivide the set of pairs obtained via the four sets of blocking criteria into five classes:

- Class 1 (1021 pairs) : Agreeing on criterion 1 and no other or simultaneously agreeing on criteria 1 and 4 and no others.
- Class 2 (624 pairs) : Agreeing on criterion 2 and no other or simultaneously agreeing on criteria 2 and 3 and no others.
- Class 3 (256 pairs) : Agreeing on criterion 3 only.
- Class 4 (244 pairs) : Agreeing on criterion 4 only.
- Class 5 (2240 pairs) : Agreeing on at least one criterion but not in classes 1-4.

Class 5 contains pairs that generally agree on two or more blocking criteria. Classes 1-5 contain 2991 matches and 1494 nonmatches and miss 59 known matches. The determination of sets of blocking criteria and classes is treated in detail in Winkler (1985a, 1987a).

We classify linkage rules by the different ways in which the individual comparison weights are computed and how resultant linkage rules are defined:

The first type, AA, of weight computation is an overall aggregate in all pairs. The second, A, is an overall aggregate in classes 1-4. The third, U, yields separate weight computations in classes 1-4. The fourth, C, adds a conditioning step that modifies the individual weight computation of Type U.

Each successive type of linkage rule involves increasingly more complex weight computations. Matches outside classes 1-5 are not considered in the results section because their number is constant for each of the four linkage rules.

Type	Individual Weight Computation	Linkage Rule
AA	Uniformly over all pairs in Classes 1-5	Over all pairs
A	Uniformly over all pairs in Classes 1-4	Designate pairs in Class 5 Links, Apply F-S to remaining pairs in Classes 1-4
U	Uniformly in each Class 1-4	Designate pairs in Class 5 Links, Apply F-S individually in Classes 1-4
C	Uniformly in each	Same as U except modify Class 1-4 weights to account for lack of independence

2.4 Evaluation Procedures

The basic evaluation technique involves comparing sizes of the region of possible links when the different types of linkage rules are applied under fixed error bounds.

As we are unable to model statistics such as the number of possible links under such complicated rules, we use the bootstrap (Efron 1987, 1982, 1979) to evaluate their distributions.

If there are sets of pairs for which the truth and falsehood of matches are known, then we can use Efron's bootstrap to estimate the variation of parameters in the following fashion:

1. Draw calibration samples of equal size with replacement.
2. Estimate individual comparison weights of the form (2.4) using the known truth and falsehood in the sample and use them to estimate total weight via (2.3).
3. Compute cutoffs LOWER and UPPER using each sample (in our application we allow at most 2 percent of the links to be nonmatches and 3 percent of the nonlinks to be matches).
4. Using individual comparison weights from step 2, compute a total comparison weight for each pair in the entire selected set of pairs. Use cutoffs from step 2 to classify pairs as links, possible links, and nonlinks.
5. Using estimates from individual samples, determine the means and variances of the cutoff weights, of the misclassification rates, and of the number of possible links.

The bounds (2 and 3 percent, step 3) are used to try to assure that the corresponding classification error rates in the entire data base are less than 5 percent.

Computations and adjustments must be performed consistently across calibration samples. Identical adjustment procedures must be used in obtaining individual adjusted weights, total weights, and cutoffs. If an individual weight is adjusted upward (step 2) by amount x or percentage y with one sample, then the same adjustment must be used with other samples.

As the underlying distributions may not be normal or may be biased and skewed, we can use new techniques of Efron (1982, 1987) to determine confidence intervals.

3. RESULTS

The results in this section comprise three parts. The first part is an overall comparison from using the four different weighting methods described in section 2.3.6. The second part contains more details about the best two methods from the first part. The third part contains results from the bootstrap evaluation.

3.1. Overall Comparison

We place fixed upper bounds of 5 percent on the number of matches misclassified as nonmatches and 2 percent on the number of nonmatches misclassified as matches. As we are using discrete data, actual error rates will generally not equal their upper bounds (Table 1, columns 2 and 3).

Table 1

Error Rates and Number of Possible Links from Applying Different Weighting Methods

Weight Type	Proportion Misclassed as		Total Classed		Possible Links
	Non-Match	Match	Non-Match	Match	
AA	.047	.020	964	2009	1512
A	.041	.015	952	2481	1052
U	.050	.020	1083	2707	695
C	.033	.019	1441	2947	97

We see that, as the complexity of the application of the weighting methodology increases, the number of possible links (size of manual review region) decreases dramatically from 1512 to 97. This indicates that the increasing complexity of the weight computations yields increasingly better decision rules.

We see that the last two methods, which both involve computing individual comparison weights separately in classes 1-4, yield the smallest sets of possible links (695 and 97, respectively).

3.2 Best Methods

We consider the best two methods, linkage rules using weights of Type U and of Type C, in greater detail. Results from applying weights of Type U and Type C are presented in Tables 2 and 3, respectively. In determining cutoff weights by class, we place rough upper bounds of 5 percent misclassified nonmatches and 2 percent misclassified matches in each class. The overall upper bound is maintained.

Table 2

Results from Using a Linkage Rule Based on Type U Weights for Delineating Matches and Nonmatches (5 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total not Classed	Total Records
	Lower	Upper	Non-Match	Match	Non-Match	Match		
1	0.5	6.5	39	14	674	264	83	1021
2	-4.5	3.5	2	4	100	115	409	624
3	-4.5	6.5	2	1	55	42	159	256
4	2.5	11.5	11	2	254	46	44	344
Totals			54	21	1083	467	695	2245

Table 3

Results from Using a Linkage Rule Based on Type C Weights for Delineating Matches and Nonmatches (3 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total not Classed	Total Records
	Lower	Upper	Non-Match	Match	Non-Match	Match		
1	4.5	7.5	28	8	692	274	55	1021
2	2.5	2.5	5	3	379	245	0	624
3	-0.5	4.5	5	6	104	110	42	256
4	8.5	8.5	9	4	266	78	0	344
Totals			47	21	1441	707	97	2245

Comparing columns 4 and 5 across tables 2 and 3, we that the corresponding numbers of misclassified matches and nonmatches are approximately the same. This is consistent with the bounding method. In every class, the linkage rule using Type C weights yields less possible links than the rule using Type U weights.

The numbers of records classified as possible links are less in classes 1 and 4 (83 versus 55 and 44 versus 0, respectively) and dramatically less in classes 2 and 3 (409 versus 0 and 159 versus 42, respectively)

One hundred percent of the pairs in classes 2 and 4 are classified by the procedure that uses Type C weights.

Two variations distinguish the linkage rule based on type C weights from the rule based on type U weights. First, we vary agreement weights associated with the four subfields of the NAME after words have been sorted by decreasing length (Table 4). The only substantial variations (greater than 2.5 on the $\log_2$ scale) occur in Class 2.

Table 4

Adjustment of Agreement Weights for Subfields Obtaind by Wordlength Sort[1]

Class	Subfield 1	2	3	4
1	.	.	−	+
2	++	++	+	+
3	+	+	−	++
4	.	+	−	+

[1] . means deviation less than 1.0,
+ , − means deviation greater than 1.0 and less than 2.5, and
++ means deviation greater than 2.5.

The second is that the agreement weight is only utilized if four corresponding subfields, the three subfields of CITY and the one STATE, agree. The variation, in effect, typically increases the **relative** distinguishing power of agreements/disagreements in subfields other than the CITY field.

The largest reduction (from 409 to 0) in the number of possible links takes place in Class 2. A slightly higher proportion (.95≈359/379) of nonlinks have an agreeing CITY field than links (.95≈223/245).

The following is an example of a match that is not designated as a link using the rule based on Type U weights but is using the rule based on Type C weights.

Name	Street	City	State	ZIP
Roberts Heat Oils	167 Sycamore St.	Dayton	OH	53315
Maxwell S Robert Heat Oil	167 Sycamore St.	Dayton	OH	53315

The first six digits of the telephone number also agreed.

The following is an example of an erroneous match using Type C weights.

Name	Street	City	State	ZIP
Molar Petro	167 Sycamore St.	Dayton	OH	53315
Petrochem	167 Sycamore St.	Dayton	OH	53315

These two companies do business from the same location and also have identical phone numbers.

The following is an example of an erroneous nonmatch using Type C weights.

Name	Street	City	State	ZIP
Johns Geo M	167 Sycamore St.	Springfield	OH	53315
Geo M Johns Jobber	167 Sycamore	Spring Field	OH	53315

Insertion or deletion of blanks in corresponding fields typically causes record pairs to be designated as a nonmatch.

3.3 Bootstrap Variation

The results of this section involve increasingly more sophisticated methods of computing bootstrap confidence intervals (Table 5). For each class, 500 replications are used in computing 90 percent confidence intervals for estimates of the number of records designated as possible links. The two error bounds are fixed at 5 percent.

Table 5

Bootstrap 90 Percent Confidence Intervals for Counts of Possible Links 500 Replications

Weight Type	Class	Ordinary Interval	BC Interval	BC_a Interval
C	1	(42,117)	(37,108)	(37,108)
C	2	(0, 0)	(7, 7)	(7, 7)
C	3	(31,154)	(34,156)	(34,156)
C	4	(0, 36)	(0, 39)	(0, 39)
U	1	(122,192)	(128,296)	(128,296)
U	2	(383,501)	(383,501)	(383,501)
U	3	(149,201)	(142,197)	(142,197)
U	4	(35, 82)	(33, 81)	(33, 81)

The first interval is the ordinary bootstrap interval that is partially based on normal theory (Efron 1979). The second interval, denoted by BC, is an interval in which a bias adjustment has been made (Efron 1979, 1982). The third interval, denoted by BC_a, is obtained using adjustments for bias and skewness (Efron 1987).

Examination of Table 5 yields that each of the intervals in respective classes are approximately the same length. If the method of adjusting to achieve weights of Type C were highly sensitive to the individual samples taken for calibration, we would expect the confidence intervals associated with Type C weights to be larger than those associated with Type U weights.

The fact that the intervals are large for either type of weight indicates the results are quite dependent on the calibrating samples. The fact that the ordinary confidence intervals are roughly the same as the BC and BC_a indicates that the respective distributions are neither biased or skewed.

The number of possible links in intervals based on Type C weights is almost always less than the corresponding intervals based on Type U weights. Only the intervals associated

with classes 3 and 4 show slight overlap. Thus, it is reasonable to accept the hypothesis that the linkage rule based on Type C weights consistently outperforms the linkage rule based on Type U weights.

4. DISCUSSION

This section is composed of three parts. The first part considers the usefulness of making comparisons that are partially dependent on other comparisons. The second contains straightforward extensions of the twostate comparisons that are utilized in this paper. The third describes methods for determining sets of blocking criteria.

4.1 Value of Dependent Comparisons

The intuitive idea of making a number of comparisons, some of which may be partially dependent on other comparisons, is that they may, when used in properly adjusted rules, yield additional distinguishing power. Newcombe and Kennedy (1962, see also Newcombe et al. 1983) have given examples of comparisons of portions of name fields that intuitively may be dependent on other comparisons. The additional comparisons, nevertheless, may yield better linkage rules than those rules that do not utilize the same additional comparisons.

The chief difficulty in using additional comparisons is properly utilizing their incremental distinguishing power. This paper's set of comparisons in particular, of subfields of the name field — is not independent in the sense of equation (2.2). The primary purpose of the set is to illustrate methods for systematically obtaining better linkage rules when the Conditional Independence Assumption is not valid.

4.2 Number and Type of Comparison States

To describe better the concepts underlying linkage rules, we restricted ourselves to consideration of twostate (e.g., agree/disagree) comparisons. Fellegi and Sunter (1969, pp. 1194-1195) give a method that extends two-state comparisons by separating the agree state into agree/take-value states.

The typical examples (Newcombe and Kennedy 1962, Newcombe et al. 1983) involve assigning agree/take-value weights that are inverses of the relative frequency of the occurrence of a given character string. Surnames are typically used.

The Fellegi-Sunter method of extending to agree/take-value states also holds for the linkage rules of this paper. In practice, those comparisons considered most suitable as agree/take-value weights can be designated as the primary set. They are estimated primarily by the basic Fellegi-Sunter procedure. The remaining weights are estimated via procedures that adjust them such as given in this paper.

Additionally, string comparator metrics (Winkler, 1985a) can be used to attach a score to those comparisons having minor spelling variations. Such variations are caused by insertions, deletions, and transpositions of characters. An example is the pair ('Zabrinsky','Zabrinky'). If we were using a frequency-based weight associated with Zabrinsky, we could adjust the weight downward to account for the spelling error.

4.3 Additional Blocking Criteria

There are two conflicting goals when a set of blocking criteria is used to reduce the number of pairs in **AXB** that receive further processing. The first is the need to reduce (drastically) the number of pairs that are processed and to obtain a set in which linkage rules can accurately delineate matches and nonmatches. The second is to obtain a set that contains as many matches from **M** as possible.

To determine whether it is feasible to look for additional sets of blocking criteria, it is first necessary to find estimates of the number of matches missed by a given set of blocking criteria. If the estimates are acceptably small, then it is not necessary to look for additional criteria.

To estimate the number of matches missed by a given set of blocking criteria, Scheuren (1983) suggested using standard capture-recapture techniques such as given in Bishop, Fienberg, and Holland (1975, Chapter 6).

Winkler (1987a) applied the techniques to the same empirical data and four sets of blocking criteria as in this paper. Using the best fitting loglinear model for the table of counts of records captured and not captured by the blocking criteria, the 95 percent confidence interval (27,160) was computed.

With any given set of pairs (in particular, those obtained by blocking) we can use Method II of Fellegi and Sunter to estimate the number of matches within the set of pairs and the agreement proportions $P(\tau \varepsilon B_i | M)$, and $P(\tau \varepsilon B_i | U)$, $i = 1, 2$, and 3, where B_i, $i = 1, 2$, and 3, are any three agreement events. The validity of the computation depends upon the Conditional Independence Assumption, assuming the number of matches within the set of pairs is greater than zero, and assuming $P(\tau \varepsilon B_i | M) \neq P(\tau \varepsilon B_i | U)$, $i = 1$, 2, and 3.

Even if the Conditional Independence Assumption is not valid, it is possible to obtain a rough estimate of the proportional number of matches within the new set of pairs. Winkler (1978b) uses the EM Algorithm to provide estimates under a weaker assumption.

If the proportion is too small or if linkage rules based on the information within the new set of pairs will not allow sufficiently accurate delineation of the matches, then it may be necessary to look for other blocking criteria or to accept a fixed number of matches missed by the given set of blocking criteria.

5. SUMMARY

The results of this paper indicate that, to reduce the size of the region of possible links when the Conditional Independence Assumption does not hold, we need to adjust estimates of individual comparison weights so that their sum yields better linkage rules.

REFERENCES

Arrelano, M. (1985), "An Implementation of the Two-Population Fellegi-Sunter Probability Linkage Model," in *Record Linkage Techniques-1985*, edited by W. Alvey and B. Kilss, U.S. Internal Revenue Service, Publication 1299 (2-86), 255-258.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975), *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA.

Cochran, W.G., and Cox, G.M. (1957), *Experimental Designs*, J. Wiley and Sons, New York.

Coulter, R.W. (1977), "An Application of a Theory for Record Linkage," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Coulter, R.W., and Mergerson, J.W. (1977), "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *Ann. Stat.*, 7, 1-26.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Methods*, SIAM, Philadelphia, PA.

Efron, B. (1987), "Better Bootstrap Confidence Intervals (with discussion)," *JASA, 82*, 171-185.

Fellegi, I.P., and Sunter, A.B. (1969), "A Theory for Record Linkage," *JASA, 40*, 1183-1210.

Kelley, R.P. (1985a), "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," in *Record Linkage Techniques-1985*, edited by W. Alvey and B. Kilss, U.S. Internal Revenue Service, Publication 1299 (2-86), 199-203.

Kelley, R.P. (1985b), "Bayesian Adjustment of the Matching Discriminant Function," paper presented at the ASA Annual Meeting in Las Vegas, NV.

Kelley, R.P. (1986), "Robustness of the Census Bureau's Record Linkage System," *ASA 1986 Proceedings of the Section on Survey Research Methods*, 620-624.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," *Science, 130*, 954-959.

Newcombe, H.B., and Kennedy, J.M. (1962), "Record Linkage," *Communications of the ACM*, 5, 563-566.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J., Strugnell, A., and Abbatt, J.D. (1983), "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers," *Comput. Biol. Med., 13*, 157-169.

Rogot, E., Schwartz, S., O'Conor, K., and Olsen, C. (1983), "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," *ASA 1983 Proceedings of the Section on Survey Research Methods*, 319-324.

Scheuren, F. (1983), "Design and Estimation for Large Federal Surveys using Administrative Records," *ASA 1983 Proceedings of the Section on Survey Research Methods*, 377-381.

Scheuren, F. (1985), "Methodological Issues in Linkage of Multiple Data Bases," in *Record Linkage Techniques 1985*, edited by W. Alvey and B. Kilss, U.S. Internal Revenue Service, Publication 1299 (2-86),155-167.

Smith, M., and Silins, J. (1981), "Generalized Iterative Record Linkage System," *ASA 1981 Proceedings of the Social Statistics Section*, 128-137.

Smith, M., Newcombe, H.B., and Dewar, R. (1983), "Automated Nationwide Death Clearance of Provincial Cancer Registry Files--The Alberta Cancer Registry Study," *ASA 1983 Proceedings of the Section on Survey Research Methods*, 300-305.

Statistics Canada/Systems Development Division (1982), "Record Linkage Software."

Statistics Canada/EDP Planning and Support Division (1984), "Record Linkage Software."

Tepping, B.J. (1968), "A Model for Optimum Linkage of Records," *JASA 63*, 1321-1332.

U. S. Department of Agriculture/Statistical Reporting Service (1979), "List Frame Development: Procedures and Software."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978a), "UNIMATCH: A Record Linkage System."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978b), "ZIPSTAN: Generalized Address Standardizer."

Winkler, W.E. (1985a), "Preprocessing of Lists and String Comparison," in Record Linkage Techniques-1985, edited by W. Alvey and B. Kilss, U.S. Internal Revenue Service, Publication 1299 (2-86), 181-187.

Winkler, W.E. (1985b), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, and Information Theory," in *Record Linkage Techniques-1985*, edited by W. Alvey and B. Kilss, U.S. Internal Revenue Service, Publication 1299 (2-86), 227-241.

Winkler, W.E. (1985c), "Exact Matching Lists of Businesses," *ASA 985 Proceedings of the Section on Survey Research Methods*, 438-443.

Winkler, W.E. (1987a), "An Application of the Fellegi-Sunter Model of Record Linkage to Lists of Businesses," Energy Information Administration Technical Report.

Winkler, W.E. (1987b), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," Technical Report.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

CONCEPTS AND PRACTICES THAT IMPROVE PROBABILISTIC LINKAGE

HOWARD B. NEWCOMBE, MARTHA E. FAIR, and PIERRE LALONDE[1]

ABSTRACT

Automated methods for probabilistic linkage of records pertaining to individual people and other entities, are improved by consciously adopting the flexibility and strategems of a perceptive human mind performing the same task. Identifiers of unlike sorts may be compared, agreements may sometimes argue against linkage, and multiple levels and values of outcome may have to be recognized to exploit discriminating power. Large files of randomly matched **unlinkable** pairs of records may be created for direct comparison with the **linked** pairs, these being analagous to the memories of a human after doing the linkages manually. The two files permit odds to be derived directly for any comparison outcome, as a ratio of frequencies among **linked** versus **unlinkable** pairs. The empiricism bypasses calculations from theory, which become laborious and error prone where the outcome definitions are complex. The simplicity fosters intuitive insights for designing efficient comparisons and uncovering trouble in existing procedures. Examples are given.

1. INTRODUCTION

Computerized record linkage is commonly thought of in one of two possible ways,

a) as an attempt to imitate **intuitively** the working of the human mind when it is carrying out the same task, or

b) as a model-based **mathematical** activity with precisely defined rules and options.

These are just complementary views of the same thing, so the arithmetic doesn't have to differ. However, the disparity in emphasis will often result in different options being chosen.

If one is to mimic the human mind one must spend a seemingly excessive amount of time observing how it works. Linkages are first done manually to acquire insights. Automated procedures are designed to reflect this experience. These have to be tested, but not just to detect errors; equally important, one needs to see what clues the machine

[1] Howard B. Newcombe is a Consultant, and was formerly Head of the Population Research Branch, Atomic Energy of Canada Limited, Chalk River; his present address is P.O. Box 135, Deep River, Ontario K0J 1P0. Martha E. Fair is Head, and Pierre Lalonde is Project Manager, Occupational and Environmental Health Research Unit, Vital Statistics and Disease Registries Section, Health Division, Statistics Canada, R.H. Coats Building, 18th Floor, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

has still missed. If a human can spot them, the machine can often be further instructed to do likewise. The practical consequences of this preoccupation, and it is indeed a preoccupation, are substantial.

We will consider three major features of any sophisticated and convenient linkage operation. They involve:

1. The use of **multi-outcome comparisons** and **comparisons between identifiers not of the same kind,** to better exploit the available discriminating power. Both have long been routine at Statistics Canada.
2. The use of **real files of randomly matched UNLINKABLE pairs,** otherwise comparable with the LINKED pairs, to provide directly measured **odds** (i.e. outcome frequencies in LINKED versus UNLINKABLE pairs, known as **frequency ratios,** the logarithms of which are called **weights).** This is a relatively recent innovation at Statistics Canada.
3. A correct procedure for **adjusting** the odds to reflect the higher or lower discriminating powers of rare versus common names and such when they **agree** or **partially agree.** This is a new development, not yet in actual use, which is needed because current methods introduce major errors when applied to the **partial agreements.**

These developments have grown out of a continuing attempt to imitate the human mind. The logic they employ should therefore be intuitively obvious, the arithmetic should be simple, and the resulting precision should reflect that of a perceptive human searcher.

2. THE CONCEPTS

All linkages, whether done by man or machine, involve comparisons of identifiers (names, birth dates, birth places, etc.) to see if the outcomes (agreements, similarities, dissimilarities, and such) are more typical of LINKED pairs, or of fortuitously matched UNLINKABLE pairs (see Table 1). The resulting **frequency ratios** are rather like the **odds** in horse racing. Or, to be more precise, they are the factors by which the overall **odds** are altered by the individual pieces of evidence.

Table 1

Linkage Concepts: Odds = Frequency Ratio

For Any given Outcome, Defined In Any Way

$$\text{Odds} = \frac{\text{Frequency in Linked Pairs}}{\text{Frequency in Unlinkable Pairs}}$$

Application: Used by man and machine;
Denominator from theory or observed

It is the logic by which this intuitively obvious concept gets used that will concern us here. There are some popular over-simplifications which obscure its true flexibility and power, and which should be corrected. For example:

- Even identifiers of **different kinds** can be compared.
- The simple **agreement** and **disagreement** outcomes often-make very poor use of the discriminating power.

- **Agreements** don't always argue for linkage, and **disagreements** don't always argue against it.
- The arithmetic doesn't have to be complicated to ensure precision of the calculated odds.

Indeed, precision depends upon a surprisingly flexible kind of reasoning. Some examples will be given:

2.1 Comparing Different Identifiers (Table 2)

When linking birth records back to the parental marriage records, the BIRTH ORDER of a child may be compared with the DURATION OF THE MARRIAGE at the time of the birth. Low birth orders are common in the early years, and a first child late in marriage is rare enough to be considered "noteworthy". Conversely, a fifth child born in the first year of marriage would be distinctly "unfashionable", unless of course it is a second marriage. Such "relatedness" between identifiers that are not of the same kind is a source of discriminating power and should be put to use.

Table 2

Linkage Concepts: Comparing Identifiers **Not of the Same Kind**
Birth Record **Versus** Parental Marriage Record

Birth Order	Duration of Marriage (YRS) 1	2	3	4	5	6+
1	+					X
2		+				
3			+			
4						
5						
6+	X					

+ = Common
X = Less Common

Agreement / disagreement has little meaning

2.2 The Matrix of Outcome "Levels" And "Values"

In the above example there is a matrix of possible combinations of BIRTH ORDERS with DURATIONS. The human mind does not memorize the whole matrix, but it is clearly aware of those combinations that are "common" and "normal", and those that are "unlikely" or even "newsworthy". The idea of **agreement** and **disagreement** has no particular use in this context; instead, it is the degree of empirical **relatedness** or **unrelatedness** that must be measured, and the frequency ratio (in LINKED versus UNLINKABLE pairs) does this.

2.3 Interpreting the Comparison Outcomes (Table 3)

It is best **not** to assume in advance which outcomes will argue **for** linkage, and which will argue **against**. For example, when linking birth records into sibship groupings it is useful to compare the birth orders on the two records. Here, **agreements** will argue **against** linkage rather than **for** it, because there shouldn't be two 1st children in the same sibship, or two 2nd children, and so on. Indeed, pre-judgments are quite unnecessary

because the correct interpretation will become obvious as soon as there are enough data to calculate a frequency ratio. In a sense, the machine finds out for itself as its "experience" increases.

Table 3

Linkage Concepts: Agreement Sometimes **Argues Against Linkage**
Birth Record **versus** Sibling Birth Record

Birth Order	Birth Order of Sibling 1	2	3	4	5	6+
1	X					
2		X				
3			X			
4				X		
5					X	
6+						

X = Unlikely (Agreement Argues **Against** Linkage)

2.4 Frequency Ratios May Be Obtained Empirically

The frequency ratios or odds, for any given outcome, can be measured directly using real files of LINKED and of UNLINKABLE pairs of records. The calculation remains simple no matter how complex the comparison procedure, provided both of these files are available. The human mind obviously does something very similar when it retains a picture of what the record pairs it LINKED were like, and how they differed from those it rejected as UNLINKABLE. These two sorts of memory probably form the empirical basis for many human quantitative judgments. The same approach may be employed by computers; these have the advantage of better memories than people, and they can be more systematic. Because this empiricism simplifies both the logic and the calculations, it is now employed routinely at Statistics Canada.

2.5 How Many Levels of Outcome? (Table 4)

For most comparisons there will be many possible levels and values of outcome, some of which have to be pooled in the interests of simplicity. However, levels with widely different **frequency ratios** (or **odds**) should not be combined because this wastes discriminating power. For example, comparisons of MONTH OF BIRTH yield 144 possible combinations or outcome "values", and these may be grouped into a spectrum of 12 "levels" (i.e. with discrepancies of 0, 1, 2, 3, ..., 11 months). Usually, the frequency ratios for discrepancies of 1, 2, 3 and even 4 months will form a graded series intermediate between those for **full agreement** at one end of the spectrum, and those for more **extreme disagreement** at the other. A simple rule of thumb is not to combine any levels where the frequency ratios differ by a factor of 2 or more. However, for discrepancies as large as 5 months and over, the frequency ratios (or odds) will usually be so similar that any loss of discriminating power due to pooling is trivial. The principle applies to any spectrum of outcome levels or values, from the comparison of any identifier. Only empirical tests will show which levels may safely be combined, and which are best kept separate.

One does not have to prove that the human mind operates in the ways one thinks it does. What is important is that the insights gained by doing the linkages oneself be shown to effectively increase the precision of the calculated odds. The improvement can usually be measured by comparing the frequency ratios from a refined comparison procedure with those from a cruder version. The interpretation should be obvious.

Table 4

Linkage Concepts: **Some Outcomes can be Combined**
Month of Birth: No. of Months Discrepant

		Month on Record B											
		J	F	M	A	M	J	J	A	S	O	N	D
	J	0	1	2	3	4	5	6	7	8	9	10	11
	F	1	0	1	2	3	4	5	6	7	8	9	10
	M	2	1	0	1	2	3	4	5	6	7	8	9
	A	3	2	1	0	1	2	3	4	5	6	7	8
Month	M	4	3	2	1	0	1	2	3	4	5	6	7
on	J	5	4	3	2	1	0	1	2	3	4	5	6
Record	J	6	5	4	3	2	1	0	1	2	3	4	5
A	A	7	6	5	4	3	2	1	0	1	2	3	4
	S	8	7	6	5	4	3	2	1	0	1	2	3
	O	9	8	7	6	5	4	3	2	1	0	1	2
	N	10	9	8	7	6	5	4	3	2	1	0	1
	D	11	10	9	8	7	6	5	4	3	2	1	0

Conclusion: don't pool levels unless frequency ratios are similar.

3. THE RELEVANCE OF MULTIPLE OUTCOMES

For virtually all identifier comparisons, recognition of a multiplicity of outcomes is the only way full use can be made of the available discriminating power. The identifiers on a death record will illustrate the point (see Table 5).

Table 5

Relevance of Multiple Outcomes: **Death Record Identifiers**

Names:	Decedent; Father; Mother; Spouse; Informant
Places:	Birth of Decedent, Father, Mother; Residence; Death
Dates:	Birth; Death
Personal:	Sex; Marital Status; Ethnic Origin; Occupation; Cause of Death

Conclusion: all comparisons can have multiple outcomes

Differing degrees of **similarity** or **relatedness** will be observed for SURNAMES, GIVEN NAMES, DATES OF BIRTH, and for pairs of INITIALS because these are sometimes inverted. PLACE OF RESIDENCE, PLACE OF WORK, and MARITAL STATUS are prone to change in non-random ways during one's life, so the interval between two records can be important in the breakdown. For PLACE OF BIRTH, which does not change, the reporting will sometimes be less than specific, and some discrepancies are in reality partial agreements as when neighbouring countries or provinces are confused with each other. Even SEX, which might be viewed as the one exception, has four possible combinations of values (MM, FF, MF, FM).

In short, multiple outcomes bring discriminating power into the open where it can be used.

4. EXAMPLES OF MULTIPLE OUTCOMES

There is no real limit to the number of possible outcomes one may recognize from a given comparison. However, in the past a practical limit was imposed by the labour of predicting from theory what the random frequencies would be in a hypothetical file of UNLINKABLE pairs of records. This constraint no longer applies at Statistics Canada, because convenient methods have been developed by one of us (P.L.) for creating and using real files of randomly matched UNLINKABLE pairs routinely. Thus, the denominators of the frequency ratios can now be observed directly.

The reasons for recognizing particular levels of outcome are usually obvious. For MONTH OF BIRTH and DAY OF BIRTH the breakdown will be by the magnitudes of the discrepancies (e.g. 0, 1, 2, 3, 4, and 5+ months or days). For YEAR OF BIRTH, age groups may be important as well. For MARITAL STATUS there should be at least nine levels representing all possible combinations of **never married, married,** and **was married,** because of the high negative discriminating powers of the unlikely outcomes (e.g. **married** followed by **never married**). Whenever one is in doubt concerning a reasonable balance between useful refinement and unwanted complexity, empirical tests are invaluable.

NAMES and GEOGRAPHIC IDENTIFIERS merit special attention. This is because they contain a major part of the total discriminating power, and because there are so many forms which **similarity** and **relatedness** can take. There are many possible comparison schemes, and the two examples are not necessarily the most efficient (see Tables 6 and 7).

Table 6

Example of a multi-outcome comparison:
given names (compared only if the initial agrees)

Full Agreements 2-7 char, all agree	(SAMUEL - SAMUEL)
Truncations 2, 3, 4-6 agree + truncation	(SAMUEL - SAM)
Agree + Disagree 1, 2, 3, 4-6 agree + disagreement	(SAMUEL - SAMPSON)

If better schemes are to be developed, they will be influenced by observing what clues one's own mind finds useful when doing the difficult linkages manually. Any new procedure will need to be tested of course, and to successfully pass the test it must recognize degrees of **relatedness** that would be missed by the cruder approach it is to replace.

This empiricism keeps the machine logic simple and easy to visualize. Here, there is an obvious parallel with the pragmatism of the human mind when it uses its' memories of the pairs it has judged to be LINKED and to be UNLINKABLE.

Table 7

Example of a Multi-outcome Comparison:
Place of Work Versus Place of Death

Died:	same city or place
	other place with same industry
	same province
	other province - western canada
	other province - eastern canada
	other country

(a finer breakdown is possible)

5. AN EXAMPLE OF EMPIRICALLY DERIVED FREQUENCY RATIOS

A single example will illustrate the use of a real file of UNLINKABLE pairs. This has to do with the FIRST and SECOND INITIALS, comparisons of which can be more complex and more important than one might suppose:

1. A **missing SECOND INITIAL** on both members of a record pair carries positive discriminating power which ought to be used.
2. **Disagreements of FIRST and SECOND INITIALS** tend to be correlated, and the resulting bias ought to be excluded.
3. **Inversions of the two INITIALS** are common and require cross comparisons to be detected. So also do **frame shifts,** as when the first of three given names is used irregularly.
4. **FIRST and SECOND INITIALS that agree or cross-agree may be used as pointers** to indicate whether the remainders of the given names should be straight-compared or cross-compared.

Earlier procedures did allow for cross comparisons, but there was no correct way to calculate the frequency ratios where **straight disagreements** were followed by **cross agreements.** (A "hit" preceded by a "miss" is not the same thing as a "hit" on first try, and should not be treated as such.)

The tidy solution is to straight-compare and cross-compare simultaneously. There are 22 possible outcomes from this (if one ignores records lacking both INITIALS and moves solitary SECOND INITIALS into first positions) and many of these outcomes can be grouped. Their random frequencies might be derived from theory, but it is simpler to get them directly from a real file of UNLINKABLE pairs.

Examples of the frequency ratios for the simultaneous straight and cross comparisons are given in Table 8, and the errors thus avoided are shown in Table 9.

Table 8
Multi-outcome Comparisons: 1st and 2nd Initials

Outcome	Frequency Ratio Links / Non-L			=	Odds		
	%						
Both 2nd Initials Present							
2 Agreements	34.60	/	.13	=	266	/	1
2 Cross-Agreements	2.06	/	.10	=	21	/	1
1 Agreement, 1 Disagreement	1.51	/	3.48	=	1	/	2
1 Cross-Agreement	.39	/	3.00	=	1	/	8
No Agreement	.07	/	19.10	=	1	/	273
One 2nd Initial Present							
1 Agreement	21.30	/	3.51	=	6	/	1
1 Cross-Agreement	2.01	/	3.17	=	1	/	2
No Agreement	.30	/	43.66	=	1	/	146
No 2nd Initial Present							
1 Agreement	37.15	/	1.69	=	22	/	1
No Agreement	.61	/	22.16	=	1	/	36
Combined	100.00		100.00				

Table 9

Multi-outcome Comparisons: 1st and 2nd Initials

True Odds (initials compared together)
versus
False Odds (initials compared separately)

	Error Factor (True / False Odds)
True Correlation Effect Exploited	
2nd Init Present on Both Records	1.5 / 1
2nd Init Absent from Both Records	1.6 / 1
2nd Init Present on one - Absent From Other	1 / 2.1
False Correlation Effect Avoided	
1st and 2nd Both Agree	1 / 1.1
1st and 2nd Both Disagree	9.1 / 1
One Agrees, Other Disagrees	1 / 4.0
False Combined Odds Avoided	
Double Cross Agreement (e.g. JW-WJ)	1 / 12.5

The procedure makes better use of discriminating power in the following ways:

a) A genuine **3-fold** spread in the odds, due to **correlated** presence or absence of the SECOND INITIAL, is put to use by the multi-outcome comparison procedure.
b) Bias from a spurious **36-fold** spread of the calculated odds, attributable to **correlated disagreements** of the INITIALS, is avoided by concatenating them.
c) A spurious **12-fold** elevation of the calculated odds for **double cross-agreements** of the INITIALS, when they are compared and cross-compared separately in a series of steps, is likewise avoided by concatenating them.

It is the real files of UNLINKABLE pairs that make this sort of sophistication both simple and convenient in practice.

6. MAKING THE ODDS VALUE-SPECIFIC (CORRECT PROCEDURE)

The **odds** or **frequency ratios** discussed so far, fail to take account of the particular discriminating powers of rare versus common names, initials, etc. A conversion is therefore needed to make them value specific. Unfortunately, the method commonly used is now known to substantially distort the odds when applied to **partial agreements.** So, a new procedure has been devised that is appropriate for both **full agreements** and **partial agreements.** It is best described by referring again to the FREQUENCY RATIO formula shown earlier (see Table 1).

Where the outcome definitions used in the formula fail to specify a particular name (or part of a name), the odds will relate to any name with an "average" frequency, i.e. neither particularly rare nor particularly common. An **adjustment** is therefore appropriate, **upward** if one wants the odds to relate to a rare name, and **downward** if one wants them to relate to a common name. Indeed, it should reflect the degree to which the particular name (or the agreement portion of it) is more rare, or less rare, than an "average" name.

No part of the "global" frequency ratio is discarded, because the information it contains is irreplaceable. (We will refer to these value non-specific frequency ratios as "global" because they embrace the whole of the outcome definition, including both the **agreement** and any **non-agreement** components.) Rather than discard anything, the frequency ratios are simply multiplied by the appropriate **adjustment factor.**

The details are simple. The "average" or **general frequency,** e.g. for male given names, is a weighted mean of the frequencies of all values. It is usually based on the file being searched, and is taken to be the sum of the squares of the individual frequencies. **General frequencies** are obtained separately for the first 1, 2, 3, 4, and 7 characters. The **ADJUSTMENT FACTOR** for the agreement portion of a particular name will be equal to its **general frequency** divided by its **specific frequency.** This ADJUSTMENT will be greater than unity for rare names, initials and such, and less than unity for the common values, just as one would expect (see Table 10).

Table 10

Examples of Specific Adjustment Factors

Value	General Frequency	Specific Frequency	Adjust Factor (Gen/Specif)
A (initial)	1 / 10	1 / 9.2	1 / 1.1
Amos (within initial)	1 / 11.5	1 / 249.0	21.7 / 1
Combined (whole name)	1 / 115	1 / 2280.8	19.8 / 1
J (initial)	1 / 10	1 / 6.4	1 / 1.6
John (within initial)	1 / 11.5	1 / 2.9	1 / 4.0
Combined (whole name)	1 / 115	1 / 18.6	1 / 6.2

Conceptually, a value specific frequency ratio could be viewed as based solely on record pairs in which the search record carries the specified value for the agreement portion of the identifier. In principle, the denominators could then be derived from a very large file of UNLINKABLE pairs. However, where neither the outcome definition (e.g. = "agreement of 4 characters followed by a relative truncation", as in ALEX - ALEXANDER) nor the particular value (e.g. agreement portion of the name = ALEX) are common, an excessively large file of UNLINKABLE pairs would be required. Certainly, files of 50,000 pairs such as are used at Statistics Canada would be far too small.

Thus, a correct method for converting **global frequency** ratios (and the logarithms of these, called **global weights**) to their value specific counterparts, will be needed for any computer linkage operation that emphasizes multiple levels of outcome.

Although correct in a general sense, the ADJUSTMENT FACTORS are capable of refinement as applied to particular situations. The proposed procedure, described here, uses the general and the specific frequencies (and their weights) precisely as they are being calculated for current practice. I.e. the distinction between names of a given length and names truncated to that length is ignored in the interest of simplicity. Refinement of the ADJUSTMENT FACTORS is therefore possible but would require additional look-up tables.

7. THE FAULTY CONVERSION PROCEDURE

The faulty procedure for converting global odds to their value specific counterparts is not just of local interest. Record linkers anywhere are likely to run into the same logical

problem. A shortcut which was tolerable for the **full agreements** had been wrongly assumed adequate for the **partial agreements.**

With the shortcut method, the **denominator of a global frequency ratio** (i.e. the outcome frequency in UNLINKABLE pairs) is replaced by a **value specific frequency** from the file being searched (the logarithm of which is called a **frequency weight**). But the discarded frequency relates to **paired records** while its replacement relates to **unpaired records,** so they are hardly equivalent even when numerically equal as in the **full agreements.**

The reason the shortcut fails with the **partial agreements** is that the outcome definitions include both an **agreement** and a **non-agreement** component. There is no way a frequency derived from **unpaired records** can reflect the **dis**agreement part of an outcome definition relating to **paired records.**

To see how the two procedures differ, let us apply each of them to **partial agreements** of the given names JOHN versus JOHN**NY**, and AMOS versus AMOS**E**. The global definition for both of these is "an agreement of the **first four characters** followed by a relative **truncation**" (see Table 11). The necessary data, as currently being calculated, are:

Global frequency ratio (first 4 agree) = 1.4% / 0.024% = 58.3 / 1.

General frequency (first 4 characters) = 0.87%.

Specific frequencies: Amos = 0.037%; John = 5.304%.

Table 11

Comparing the Faulty with the Correct Conversion Procedure
(using identifier pairs Amos - Amo**se**, and John - John**ny**)

Agreement Portion (plus non-agreement)	**Frequency Ratio** Links / NonL	=	**Odds**
	%		
Faulty Procedure (denominator changed to specific frequency)			
Global 4-Agr (+ Truncation)	1.4 / 0.024	=	58.3 / 1
Amos (blank - E)	1.4 / 0.037	=	37.6 / 1
John (blank - NY)	1.4 / 5.304	=	1 / 3.8
Correct Procedure (multiplying by specific adjustment factor)			
Global 4-Agr (+ truncation)	1.4 / 0.024	=	58.3 / 1
Amos (blank - E)	(times 0.87 / 0.037)	=	1370.8 / 1
John (blank NY)	(times 0.87 / 5.304)	=	9.6 / 1

Here, the faulty procedure yields odds that are too low by **36-fold.** (E.g., a rare value like AMOS should raise the odds, not lower them.) This error tends to increase with the multiplicity of the outcome levels. More particularly, it occurs wherever the **general frequency** (in this case **0.87%**) differs from the **global frequency** among UNLINKABLE pairs (0.024%), and the error is equal to the ratio of the two (0.87% /0.024% = **36.25-fold**).

This is **not** an extreme example. When various levels **of partial agreement** are considered, the faulty conversion procedure can be seen to distort the odds by anywhere from 2-fold to 2000-fold (see Table 12). Some 17 per cent of GIVEN NAME comparisons are affected in the linked pairs. (If SURNAMES are considered as well, the proportion rises to 20 per cent.)

Table 12

Magnitudes of the Errors from the Faulty Conversions
(Comparing given names of 2-7 characters; initial agrees)

Agreement Level	General Freq (%)	Global Freq (Non-L) (%)	Error Factor (Gen/Glob)	Linked Pairs Affected (%)
Full Agreement	.98 est.	.982	1.0-fold	83.1
4-6 Agr + Trunc	.87	.024	36.3-fold	1.6
3 Agr + Trunc	1.96	.014	140.0-fold	.3
2 Agr + Trunc	3.69	.0016	2306.3-fold	.1
4-6 Agr + Disag	.87	.085	10.2-fold	4.2
3 Agr + Disag	1.96	.142	13.8-fold	3.3
2 Agr + Disag	3.69	1.383	2.7-fold	2.2
1 Agr + Disag	9.85	4.395	2.2-fold	5.2

The mechanics of the two procedures are not so very different. The faulty one used look-up tables of **specific frequencies** (as logarithms called **frequency weights**) for the various values of the names, initials, and such, and for shortened versions of these. The new and correct procedure requires only that these look-up tables contain instead the **specific adjustment factors** (or their logarithms, the **specific adjustment weights**), and that the global frequency ratios be retained intact and adjusted by the appropriate factors.

In short, a correct method of converting **global frequency ratios** (and **global weights**) to their value specific counterparts, avoids the frequent major errors introduced by an older faulty procedure, and with no added cost or inconvenience. The weights themselves, and the adjustment factors, are still open to refinement but this is a separate matter.

What makes this test simple and conclusive is the use of real files of randomly matched UNLINKABLE pairs, to observe directly how often the various **partial agreement** outcomes will occur by pure chance. Indeed, the nature and the magnitude of the error was only discovered because the real files of UNLINKABLE pairs effectively bypassed the complexity of the mathematics and thereby make it easier to picture in the mind what was going wrong.

8. THE PURPOSE OF REFINEMENT

The aim of almost any refinement of a linkage procedure is necessarily modest. This is because the majority of potentially linkable records will be correctly matched no matter how crude the approach, and because there will always be others that do not get linked no matter how elegant the methods. Thus, refinements can only influence the end result where the needed discriminating power exists in a pair of records but is not readily accessible. It follows that measurements of the overall precision of any linkage operation are apt to tell more about the state of the files than about the efficiency of the methods.

Having said that, there is still good reason to improve the effectiveness of linkage wherever this can be done without undue cost. The two main sources of inefficiency are **logical errors,** and **failures to exploit the discriminating power** that is there. In the past, both have been associated with excessive emphasis on the simple **agreement/disagreement** outcomes.

Detection of logical errors and missed opportunities, has usually occurred during visual scrutiny of the more difficult links. Such reviews enable one to observe the logical steps going on in one's own mind. This kind of intuitive and partially unconscious reasoning is less influenced by artificial constraints than is the deliberately formal kind. More importantly, wherever the mind does an obviously better job than the computer, the reason is worth discovering. The secret will usually lie in some initially unconscious mental stratagem. Once its nature is revealed, the computer can then often be instructed to employ it too.

A realistic aim, therefore, is for the practice of automated linkage to evolve, as older refinements come into routine use and as further improvements get added, until it matches or surpasses the skill of a perceptive human.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

RECORD LINKAGE-METHODOLOGY AND ITS APPLICATION

MIKE EAGEN and TED HILL[1]

ABSTRACT

Record Linkage is a well established discipline. Throughout the 1970's and 1980's many refinements were introduced: improved linkage rules were developed which use available information more effectively; cross-comparison of fields has become commonplace; procedures were developed to handle fields violating data independence assumptions; alternative weighting schemes were developed; and, in Canada, a generalized computer system was developed to minimize computer system expenditures.

These developments have advanced the state-of-the-art almost to the limit of the information available from the data. Nevertheless, a continuing gap exists between the methodology available and the actual applications.

The remaining challenges for linkage experts are:

(i) to develop an integrated methodological approach(es); and,

(ii) to enhance system tools which would enable straightforward implementation of the integrated approach(es).

We propose a strategy to meet these challenges.

1. INTRODUCTION

Record Linkage typically consists of comparing records on two incoming files in order to determine if they represent the same individual (or entity). In this paper, we consider only this application type and, further, we have limited ourselves to the situation where each record is linked to zero or one record on the other file. The comparison involves executing a number of linkage rules. A linkage rule may be very straightforward (e.g., compare surnames) or more complex (compare addresses consisting of a number of fields on each file). Each linkage rule produces an outcome. Possible outcomes for the surname comparison might include:

(i) surnames agree;

(ii) first four characters and phonetically encoded name (e.g., NYSIIS, SOUNDEX, Name Search Key) agree;

(iii) coded name agrees;

[1] Mike Eagen, Goss, Gilroy & Associates Ltd., 400-222 Queen Street, Ottawa, Canada K1P 5V9. Ted Hill, Statistics Canada, 2405 Main Building, Ottawa, Canada K1A 0T6.

(iv) first character agrees;

(v) names disagree; and,

(vi) surname missing on one or both files.

The second through fourth outcomes above are generally referred to as levels of partial agreement. Linkage rules can also produce values for certain outcomes. When comparing two records with the surname SMITH, the outcome is AGREE and the value is SMITH.

In practice, it is not realistic (or necessary) to conduct all possible comparisons. To reduce costs, the linkage rules are not applied to all record pairs. Other record pairs will be rejected after only some of the linkage rules have been executed.

For the remaining record pairs, weights are assigned based on the outcomes (and values) for each linkage rule. Total weights are computed for the record pair and linkage decisions are made on this basis. (Record pairs with high weights are considered definite links; those with low weights are considered definite non-links. Usually there will be a group in the middle classified as possible links.)

The goal of the practitioner is to minimize the size of the group of possible links without increasing the risk of making incorrect decisions.

2. HISTORY OF RECORD LINKAGE

Record Linkage has a long and illustrious history. The first work was conducted in the 1950's (Newcombe, H.B., Kennedy, J.M., Axford, S.L., and James, A.P. (1959)). Throughout the 1960's, the computer became a fundamental tool in linkage applications. Also, during that decade, a theory for Record Linkage developed which culminated in the classic paper in the field (Fellegi, I.P., Sunter, A.B., (1969)). This paper provided support to the work of practitioners, and exposed a much wider audience to the techniques.

The 1970's was a period of significant growth for Record Linkage. Applications flourished, particularly in the field of epidemiology where Record Linkage became a fundamental technique in establishing environmental risk factors. The techniques made possible research which could not otherwise have been done, thereby saving lives and easing the concerns of workers and their families. To this day, epidemiology is a dominant application area for Record Linkage. Also in the 1970's, central statistical agencies such as Statistics Canada began using the technique to construct and unduplicate survey frames and central registries of businesses and farms. The growth in applications necessitated the development of powerful and flexible computer systems such as the Generalized Iterative Record Linkage System at Statistics Canada (Hill, T., Pringmill, F. (1985)). During the 1970's, many methodological improvements were made. In particular, comparison rules which created multiple outcomes were defined thereby making better use of the discriminating power of the data (Newcombe, Fair, Lalonde, 1987). Thus, Record Linkage emerged as a true discipline in the sense that it became a field dominated by specialists who, based on their experience and knowledge, could achieve substantially better results than novices.

During the 1980's, methodology improvements have continued and, in our view, have virtually peaked. Also, centres of expertise have developed which are highly skilled in the application of this refined methodology. In Canada, for instance, the Health Division of Statistics Canada and the Ontario Cancer Foundation have established units whose major mandate is to apply Record Linkage to epidemiological problems. With their vast experience, and refined tools, these organizations can produce rigorous scientific results which are fully accepted by their colleagues. Nevertheless, in most hands, Record

Linkage remains a very difficult and error-prone technique. The difficulty concerns weighting, particularly when complex linkage rules are used. More specifically, we contend that strategies and tools for weighting have lagged far behind the advances made in other areas such as greater discrimination of rules. Unfortunately, this lag makes it very difficult for any but the most skilled practitioners to realize the potential improvements promised by methodological advancements.

In this paper, we describe and assess the various approaches to weighting in some common situations; we explain the need for a cohesive rather than fragmented approach to weighting; we describe the approach we advocate (and the requirements for its consistent implementation); we identify areas for future research; and, we offer our conclusions.

3. THEORY OF WEIGHTING

The basic theory of weighting is straightforward. First of all, linkage rules must be defined in such a way that the rules are statistically independent, i.e., the probabilities of various outcomes for any one rule are independent of the outcomes of other rules. With this property, weights can be obtained for each rule independent of the others. For rule i and outcome j, the weight is given by the following probability ratio:

$$W_{ij} = \frac{\text{Prob (j/records are linked}}{\text{(Prob (j/records are not linked)}} \qquad (1)$$

It is easy to see that outcomes which are more likely among links will have weights greater than one. Conversely, an outcome which is more likely among non-links will have a weight between 0 and 1. When logarithms are taken (which is customary) these weights will be positive and negative respectively.

The total weight for any two records compared will simply be the product of the weights (1) (or the sum of the logarithms) for all linkage rules.

The problem, of course, is estimating the probabilities used to calculate the weights. Two basic approaches have been used. The first approach (suggested by Fellegi and Sunter (1969)) is to work with the theoretical formulae. The second approach, suggested by many authors including Newcombe, Fair and Lalonde (1987), involves estimating the required probabilities directly from sample files of links and non-links created for this purpose. We refer to this method as the direct method. The relative strengths and weaknesses of these approaches are best illustrated by examining various types of linkage rules.

3.1 Exact Agreement

It is easy to see that the

Prob (exact agreement on value k | records are linked)

= Prob (exact agreement | records are linked) × Prob (value k | exact agreement | records are linked)

$$= \text{Prob (exact agreement | records are linked)} \times \frac{n_L(k)}{N_L}$$

where, $n_L(k)$ = frequency of value k among the links

N_L = number of links

In the absence of errors in the data it is easily seen that

Prob (exact agreement or value k | records are not linked)

= Prob (value k on file A) × Prob (valuek on file B)

In practice, of course, the files will contain errors. However, since they work in both directions it is reasonable to assume that they effectively cancel each other and the use of $RF_A(k) \times RF_B(k)$ is reasonable.

Substitution of these results into (1) provides

$$W_k = \frac{RF_L(k) \times \text{Prop (Exact agreement among links)}}{RF_A(k) \times RF_B(k)} \tag{2}$$

where,

k represents the value agreed upon (e.g., if the rules related to comparison of the given names, then (2) gives the weight for exact agreement on value k, say Philippe);

$RF_L(k)$ is the relative frequency of the value, k, among true links and is generally not known in advance;

$RF_A(k)$ and $RF_B(k)$ are the relative frequencies on the two files being linked; and,

Prop (exact agreement among links) represents the proportion of true links which produced the outcome: "exact agreement".

Eagen (1978) has dealt with this situation in greater detail.

Conventionally, the smaller of the two files being linked is referred to as File A. In many applications (especially in the field of epidemiology), File A is much smaller than File B. Since a much greater proportion of File A records than of File B records will be linked, it follows that the frequency distribution of values for the links is best estimated from File A. In many instances, it is entirely reasonable to assume that $RF_L(k) = RF_A(k)$ producing

$$W_k = \frac{\text{Prop (exact agreement among links)}}{RF_B(k)} \tag{3}$$

It is straightforward to obtain $RF_B(k)$, for all k, from the large file. The proportion of exact agreement among links for the field in question can be estimated iteratively as one converges to the true links. This capability is provided by the GIRLS system at Statistics Canada. In general, this proportion will be near one and can safely be ignored under the following conditions:

- all linkage rules involve a single field and have outcomes of exact agreement, disagreement and missing, only; and,
- the proportion of exact agreement among true links does not vary substantially from field to field.

Under these conditions, weights for exact agreement depend only on the frequencies on File B, the large file. This is a significant result which vastly simplifies weighting. However, misapplication of the result has been common.

Direct Approach

The direct approach estimates the two probabilities in (1) by determining the proportions of cases of exact agreement for the links and non-links (ignoring the value agreed upon) and then adjusting the weights based on the frequency of the value agreed upon. The proponents of this approach have not described their method mathematically but formula (4) below accurately describes their approach.

$$W_k = \frac{\text{Prop (exact agreement among links)}}{\text{Prop (exact agreement among non-links)}} \times AF(k) \qquad (4)$$

where,

$$AF(k) = \text{Adjustment Factor for value } k$$

$$= \frac{GF_B}{RF_B(k)}$$

where,

$$GF_B = \text{General Frequency for Exact Agreement}$$

$$= \sum_k RF_B(k)^2$$

It would be more correct to use

$$GF_{AB} = \sum_k RF_A(k) \times RF_B(k)$$

but this is much more difficult in practice and is not commonly done.

The proportion of exact agreement among non-links is obtained by creating, at random, a file of comparisons from the incoming files. Since virtually all comparisons in the sample will be non-links, the approach, called "The Method of UnLinkable Pairs", is sound. The proportion of exact agreement among links, is estimated iteratively as above.

Comments

The two methods are theoretically equivalent in this case. (Strictly speaking, theoretical equivalence holds only if one uses GF_{AB} in place of GF_B in calculating $AF(k)$. However, pragmatically speaking, the two methods are essentially equivalent even when GF_B is used.) The theoretical approach has often been used in practice since it is relatively straightforward to apply to this situation. The direct approach has no advantages in weighting exact agreements and entails the additional overhead of creating and analyzing the file of unlinkable pairs.

3.2 Partial Agreement

Experienced practitioners know that even when exact agreement does not occur, some results argue more strongly for linkage than others. For example, when comparing given names, the comparisons Phil - Philip; Phil-Philippe; and, Phillip - Philippe argue more strongly for linkage than results such as Phillip - George. It is common, thus, to define partial agreement outcomes which recognize these situations. If we define agreement on the first four characters as a partial agreement, then the first three comparisons above

will be recognized as partial agreements. This is clearly a sound practice but it does make the task of weighting more difficult. While the difficulty is not extreme, errors in practice are common. Let us examine what the various approaches have to offer.

Theoretical Approach

Eagen (1978) showed that the appropriate weight formula for this situation is:

$$W_m = \frac{RF_L(m) \times \text{Prop (Partial Agreement Among Links)}}{RF_A(m) \times RF_B(m) - \sum_{k \varepsilon m} RF_A(k) \times RF_B(k)} \tag{5}$$

where m is the value partially agreed upon (e.g., Phil) k is a value containing m (e.g., Philip), and all other items are defined as previously

The summation term in the denominator is a correction for the fact that exact agreement has not occurred. (**Note:** If both files contain Phil, for example, as the full given name, the outcome will be exact agreement.)

This formula is more problematic than formula (2). First of all, the proportion in the numerator is typically not near one and must be estimated iteratively. Secondly, the formula does not simplify by cancellation whatever we assume. Thirdly, we must now assume that $RF_L = RF_A = RF_B$ if only one file is to be used in calculation. This is much stronger than assuming $RF_L = RF_A$ Fourthly, system calculation of these weights is more complex than that for exact agreement weights. Eagen (1978) provided specifications for this situation but, to the best of our knowledge, they have never been applied. In many cases the exact agreement formula has been applied even though it is known to be wrong and as Newcombe, Fair and Lalonde (1987) have shown, underestimates the correct weight (sometimes drastically).

We note that the formula above applies equally well to multiple levels of partial agreement. System calculation, however, becomes more complex with multiple levels of partial agreement.

Direct Approach

The direct approach has the major advantage that the methodology is independent of the nature of the outcome. The methodology for weighting partial agreements is identical to that for weighting full agreements.

The formula is as follows:

$$\text{Weight} = \frac{\text{Prop (partial agreement among links)}}{\text{Prop (partial agreement among non-links)}} \times AF(m) \tag{6}$$

where, $$AF(m) = \frac{GF_B}{RF_B(m)}$$

Note that in calculating the proportions, partial agreement means partial but not full agreement.

This approach produces weights which are somewhat different from the theoretical weights. It is easy to show that the weights from (6) will "on average" be the same as those from (5), although the weights for any specific value, m, would likely differ. We do not believe these differences to be important in practice but substantive research is lacking on this question.

Intuitive or Experiential Approach

Under this approach, weights for partial agreement are set somewhere between the global (or average) frequency weight for agreement and the disagreement weight. This seems very sensible and intuitively is better than using simpler linkage rules without partial agreement outcomes. However, with intuitions like ours, the approach can be problematic. The range of choice is often large and one must have some understanding of the underlying probabilities or it is possible to merely stir up the already muddied waters. Selection of weights can be particularly problematic when multiple levels of partial agreement are used. The approach is best applied in conjunction with the adjustment factors used in the direct approach.

Comments

The direct approach is clearly superior for this situation. It provides a degree of rigour that the intuitive approach cannot. Also, it is easier to implement than the theoretical approach. The major obstacle to its use is the lack of readily available tools for implementation.

3.3 More Complex Rules

Experienced practitioners have found it worthwhile to devise rules which cover a whole range of additional situations. Partial agreement situations have been split into agreement + missing (e.g., Phil - Philip) and agreement + disagreement (e.g., Phillip - Philippe) situations on the grounds that agreement + missing outcomes are more supportive of linkage. Another common tool is to conduct cross-comparisons (e.g., compare first given name on file A to second given name on file B). A third type of rule is to construct a single linkage rule involving many correlated fields (e.g., address data). This is useful since if, for instance, postal or zip code agrees, there is no extra information in the fact that the municipality agrees and it would be a mistake to assign a positive weight for agreement on municipality. On the other hand, if postal code disagrees, agreement on municipality may well provide a useful contribution to the linkage decision.

Many such rules are entirely sensible and, potentially can be very useful in making correct linkage decisions in close cases. On the other hand, if these rules are not properly weighted, they can muddle an already confusing situation. Even worse, they can take what would have been clear decisions and move them into the grey area. (Since these rules typically assign a positive weight to a situation which would otherwise not be recognized, the major danger is in giving non-links total weights equivalent to some true links.) Since resolving the grey area is the major rationale for weighting (any algorithm can resolve the easy cases), a correct weighting strategy is extremely important for these situations (even though they may occur relatively infrequently).

In general, the theoretical approach has little to offer for these situations since an adequate formulation of the underlying probabilities typically cannot be constructed from theory.

On the other hand, the direct approach has great promise. Since it simply observes the frequencies of the observed outcomes among "links" and "non-links", the complexity of the rule has no impact on the difficulty of weight derivation.

The intuitive or experiential approach is also feasible here for practitioners who are very familiar with their data. Our experience with this approach has not been rewarding and we recommend against its use.

The formula for these situations is:

$$\text{Weight} = \frac{\text{Prop (outcome among links)}}{\text{Prop (outcome among non-links)}} \times \text{AF(value)} \tag{7}$$

This is merely a generalized version of (4) and (6).

In some cases, the Adjustment Factor is dropped resulting in the simplified formula:

$$\text{Weight} = \frac{\text{Prop (outcome among links)}}{\text{Prop (outcome among non-links)}} \tag{8}$$

This formula provides what are typically called global (or average) weights. Typically, the global weights are always used in the initial linkage. Once the clear non-links are discarded, it is usually fruitful to substitute the value-specific weights ((4), (6) and (7)). However, if the possible values all have similar frequencies (e.g., month of birth) the Adjustment Factors will also be similar and the experienced practitioner will dispense with them in the interest of cost effectiveness. In other cases, involving very complex rules, it would be useful to incorporate the adjustment factors but it may be too difficult to calculate them in practice. The address rule mentioned in the first paragraph of this section is such a case.

3.4 Disagreement Weights

We have avoided any discussion of disagreement weights in the previous sections. In general, these are easily obtained. Eagen (1978) has provided formulae for the theoretical approach. Using the direct approach, disagreement is an outcome like any other with the additional simplification that adjustment factors are not required.

4. SELECTING AN OVERALL STRATEGY

The ideal linkage situation is, of course, when you have many common data items and have previous experience from a similar situation. In such circumstances, it may be entirely feasible to use only agreement-disagreement rules which are relatively easy to weight. If more complex rules are required, weights may be derived from past experience. However, life is not usually this kind.

In a more typical application, there will be some simple agreement-disagreement rules; some rules involving partial agreement; and, some more complex rules involving cross-comparison or multiple fields. It is also common (at least for us) to begin with simple rules and later realize that more complex rules could be helpful in resolving doubtful cases.

The discussion in section 2, which relates to individual linkage rules, may lead one to conclude that the choices are relatively straightforward. However, one must keep in mind that it is the total weight for a comparison that is used to make the linkage decision. As noted earlier, the total weight is the product of the individual weights (or the sum of the logarithms). Clearly, this is a "weakest link" situation (the pun is regrettable) and defining weights for one rule which are incorrect (or inconsistent with other weights) can destroy the integrity of the whole process.

Clearly, it is essential to devise an overall strategy which ensures that integrity is maintained. Unfortunately, this is very difficult. In particular, we note that the advances of recent years (i.e., refinement of linkage rules) have greatly magnified this difficulty.

We have seen linkages (indeed we have conducted linkages) where all three weighting strategies have been applied to various rules. Viewed in isolation, the weighting strategy

for each rule is sound. Also, each rule clearly uses additional information and should improve the linkage. Somehow, however, it doesn't usually add up. As the project nears completion, it becomes clear that a relatively unimportant rule makes a bigger contribution than a more important rule. The conscientious practitioner adjusts the weights only to find another pair or triplet of rules (typically including the one just adjusted) with the same problem. One can continue such adjustments. Unfortunately, convergence is typically elusive.

We believe that choosing an overall strategy for weighting is the most important, most difficult and, frequently, most ignored component of Record Linkage Methodology.

We are aware of only two strategies which avoid the dilemma described above.

Strategy 1: Keep it Simple

This strategy, which essentially consists of avoiding the leading edge, has been used with much success in many applications. It is particularly appropriate when one has many data items to compare or equivalently when data is of very high quality and not subject to change. In the absence of these circumstances, this approach is also appropriate when the cost of incorrect decisions (i.e., false links and missed links) is low.

Typically, such applications use simple rules, which involve a single field and have outcomes of agree, disagree and missing. They may use only global weights or they may use value-specific weights utilizing formula (2) or (3) (or an equivalent approach).

However, in many applications (especially in epidemiological work) it is critical that virtually all linkage decisions be correct. Often, data is limited and of doubtful quality. In these applications, complex linkage rules are unavoidable and Strategy 2 must be used.

Strategy 2: Direct Weighting for all Rules

The direct method is the only one which can be applied to any linkage rule. Consequently, for applications which demand more complex rules, it must be used for all rules if the integrity of the weighting process is to be guaranteed.

5. APPLICATION OF DIRECT WEIGHTING

Application of direct weighting is, however, far from common and has its own difficulties. Indeed, we are aware of only one organization (the Health Division at Statistics Canada) which is experienced with this approach (Newcombe, Fair, Lalonde (1987)). Further, we are aware of only one existing computer system (GIRLS, Statistics Canada (Hill T., Pringmill F. (1985)) which can accommodate this methodology. In order to make this approach more widely usable, the following developments are necessary:

- generalized computer tools must be developed to facilitate calculations of the necessary proportions and adjustment factors. This is not complex, in our view, but even at Statistics Canada, these tools are not currently available;
- since estimation of the numerator of the weight formulae (i.e., Prop. (outcome among links)) frequently must be done iteratively, the success and cost of the method is dependent on good initial estimates. Experienced practitioners must make the results of their work available and "new" rules must first be tested in an experimental situation if good initial estimates are to be available;
- in the short term, organizations conducting sophisticated linkages, can expect reliable results only by including linkage experts on their application teams;

- we recommend that organizations, such as Statistics Canada, which are active in Record Linkage establish a Record Linkage consulting function to ensure consistent and accurate implementation throughout the organization; and,
- detailed instruction manuals must be prepared which make it possible for relative novices to correctly apply this methodology. Newcombe (1986) is the best source at present. However, the emphasis is on strategy rather than implementation (this is inevitable since the necessary tools are not readily available at present).

6. FUTURE RESEARCH

There are four major areas where additional research work is required. The first two represent priority areas.

6.1 Cost/Benefit Analyses

We are unaware of any analyses of the incremental costs and benefits of the refined linkage rules which have come into use in recent years. While it is intuitively clear that more refined rules, if properly weighted, should result in improved linkage decisions it is also true that:

- most linkage decisions are easily made with any reasonable set of rules; and,
- refinements result in significant additional costs.

In order to make an intelligent choice between Strategies 1 and 2 in section 5, it is important to have some idea of the relative costs and benefits.

6.2 Weights from Other Applications

As noted in section 5, Strategy 2 involves iteration. Iterative processes are most likely to converge when one has a good starting point and the better the starting point, the fewer iterations are required (which reduces costs and improves timeliness). Since starting points frequently must be obtained from past experience, the publication of results by experienced practitioners would be very helpful to others. We have found it relatively easy (but not particularly helpful) to obtain information on the net weights (see formula (7)) used in other applications. Since a starting point is required only for the proportion of the outcome among the links, this is the result which would be most useful. It is also of benefit to obtain from these past applications the:

- proportion of the outcome among the non-links. In general, one should calculate this for each new application but it is useful to have these results from the past application as a reflection of the similarity of the two applications. Major differences may suggest the need for an alternate linkage rule; and,
- the range of the Adjustment Factors. The major value of this is that it clarifies the importance of these Factors for the particular rule.

Of course, if only the net weight (or some other breakdown of the weight) is provided, the past experience will be of limited value.

6.3 Development of New Linkage Rules

Ideally, a new linkage rule would be developed in an experimental setting where links and non-links are already known since it is relatively easy to refine the rule in such circumstances. However, creativity is spurred by adversity, and most new rules will be developed in a production environment. Publication of results, in the form described above, is important to the continuing development of Record Linkage methodology.

6.4 Sampling of Unlinkable Pairs

The usual method for sampling of unlinkable pairs, developed by Pierre Lalonde of Statistics Canada (Newcombe, Fair, Lalonde (1987)) involves selecting a random sample from all possible comparisons. Since the number of possible comparisons is the product of the number of records on the two files and the number of actual links is at most the number of records on the smaller file, it follows that the probability of including true links in the sample is very small. This approach is entirely sound.

On reflection, however, one realizes that most possible comparisons never reach the weighting stage since they are so clearly non-links. These comparisons are either never done (due to blocking strategies) or are dismissed early on by initial linkage algorithms. This raises the possibility that perhaps only those non-links which meet some minimal standard of commonality should be used in weight calculation. Since the purpose of weighting is to distinguish between actual links and those non-links which, in some sense, appear to be links, perhaps the appropriate global weight is given by:

$$\text{Weight} = \frac{\text{Prop (outcome among links)}}{\text{Prop (outcome among "close" non-links)}}$$

This appears to be a promising approach although preliminary research by the Health Division of Statistics Canada was discouraging.

7. CONCLUSIONS

Our conclusions are as follows:

- the refinement of linkage rules, while clearly a positive development, has complicated the task of weighting and necessitates the use of an integrated methodology which guarantees the integrity of the linkage decisions;
- the only feasible methodology is difficult to apply and is currently not available to most practitioners. This can be addressed only by development of new generalized system tools;
- in the short-term, significant expertise is required for sophisticated linkages. This necessitates the establishment of a formal Record Linkage consulting function; and,
- the current priority for research is in the application area rather than in the development of new methodologies since current applications often do not approach the potential provided by existing methodology. Special care is required to ensure that information from past experience is made available in a useful format.

REFERENCES

Eagen, M. (1978). "Specification for the Calculation of Weights for GIRLS". Technical Report. Institutional and Agriculture Survey Methods Division, Statistics Canada.

Fellegi, I.P., and Sunter, A.B. (1969), "A Theory for Record Linkage". Journal of the American Statistical Association, 1182-1210.

Hill, T. and Pringmill, F. (1985). "A Generalized Iterative Record Linkage System". Proceedings of the Workshop on Exact Matching Methodologies. Arlington, Virginia, May 9-10, 1985.

Newcombe, H.B. (1986). "Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration and Business". Publication pending by Oxford University Press.

Newcombe, H.B., Fair, M.E., and Lalonde, P. (1987). "Concepts and Practices that Improve Probabilistic Linkage". International Symposium on STATISTICAL USES OF ADMINISTRATIVE DATA, sponsored by Statistics Canada, Ottawa, Canada, 1987 November 23-25.

SESSION III: INVITED PAPERS

INTEGRATED APPROACHES TO DATA DEVELOPMENT

Chairperson: B. Petrie, Statistics Canada

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

USE OF ADMINISTRATIVE DATA IN THE BUSINESS SURVEY REDESIGN PROJECT

MICHAEL COLLEDGE[1]

ABSTRACT

The Business Survey Redesign Project currently in progress at Statistics Canada is a major undertaking which will dramatically affect the economic statistics program. At the heart of the project is the redesign of concepts, procedures and systems for the provision of survey frames and for the use of income tax data. In addition, the project presents an opportunity to review the whole suite of economic surveys and to develop generic procedures and systems, with the objectives of facilitating data integration and reducing costs and respondent burden. Exploitation of administrative data, both as a basis for building survey frames and as a replacement for direct data collection, is a crucial part of the approach. This paper provides a summary of the project strategy and its implementation, with particular reference to administrative data.

1. INTRODUCTION

The paper is in five parts. This first section contains a brief overview of the Business Survey Redesign Project, covering the objectives, the strategy for achieving them, and the relevant administrative sources. The following three sections focus on the principal applications of administrative data within the Project. Section 2 deals with the use of these data for business survey frame definition and maintenance. Section 3 describes the partial replacement of annual survey data collection by procedures based on income tax returns. Section 4 refers to applications of administrative data being introduced in the redesign of other survey functions, and indicates the effects being made in this context to develop generalized methods and systems.

The inherent limitations of administrative data, arising from their focus on administrative rather than statistical needs, are exemplified, and the procedures introduced to cope with these limitations are described. Policy issues, for example concerning data exchange, security and confidentiality not covered in the paper. For more details in this area see, for example, Brackstone (1987). Quality issues are mentioned but not discussed in detail (see Lussier and Colledge, 1987). The paper concludes with a summary and some general remarks regarding the significant role which administrative data have in business surveys.

[1] Michael Colledge, Assistant Director, Business Survey Methods Division, Statistics Canada, 11th Floor Section A, R.H. Coats Building, Ottawa, Ontario. K1A 0T6.

1.1 Business Survey Redesign Project

The economic statistics program at Statistics Canada includes surveys of financial, industrial, commodity, employment, capital expenditure and taxation statistics collected on a monthly, quarterly, annual and occasional basis. There are roughly 300 surveys (depending upon the precise definition of a survey), of which about 125 are subannual and the balance annual or occasional.

In 1985, Statistics Canada initiated the Business Survey Redesign Project (BSRP), the main objective of which is to standardize and integrate all the economic program systems and output data. This goal is to be achieved through the mandatory use of a common central frame and of a generalized survey methodology, which should not only facilitate the integration of statistics and rationalize operations, but also reduce the cost of future survey designs, redesigns and upgrades as specific methodologies and tailormade systems are replaced by generalized ones.

It should be mentioned at this point that the term "business", which appears within the project title, is used within Statistics Canada to include units of economic production engaged not only in the trade and commercial service industries but also in manufacturing, construction, transportation, and professional activities, etc.

The BSRP constitutes a major effort to build quality into the economic program. The benefits which it is hoped will materialize include:

(a) standardization of concepts, definitions, classification schemes and survey procedures;

(b) more comprehensive frame coverage, more precise delineation of large economic units and their reporting arrangements, and more reliable industrial classification;

(c) increased use of administrative data, reduced response burden and improved respondent relations;

(d) reduction in overall frame maintenance, mailout and data collection costs;

(e) enhanced facilities for integrating data and increased scope for audit;

(f) development of generalized systems for a wide range of survey functions including automated or computer assisted industrial coding, sample size determination, allocation and selection, edit, imputation, etc.

The strategy for achieving these objectives contains two important themes. The first theme is standardization: the introduction of standardized concepts, generic methods and generalized systems, and, in particular, a new central frame data base. The second theme is the use of administrative data, both for survey frame definition and maintenance, and to replace direct data collection. Thus administrative data are a cornerstone of the BSRP. Full details of the objectives and strategy are given by Cain et al (1984) and have been summarized by Colledge and Lussier (1985, 1987) and Colledge (1987).

1.2 Sources of Administrative Data

There are many divers administrative sources, both government (federal, provincial and municipal) and commercial. Examples are: payroll deductions, corporate income tax, personal income tax, unemployment insurance, imports, exports, chartered banks, registered trust companies, airlines, vessels, transit systems, utilities (electricity, water, telephone), trade associations, marketing boards, etc. Early in the BSRP all potential sources were listed and evaluated in terms of their utility for business surveys (Bankier et al, 1985). It was decided to focus the redesign on three main sources, all from Revenue Canada, Taxation, namely payroll deduction, corporate income tax and personal income tax. (If a business transfer/value added tax were to come into existence it would form a very important fourth source. For the moment the plans for its introduction are so

tentative as to preclude its incorporation within the strategy.) In the following sections the role of these sources will be elaborated.

2. ADMINISTRATIVE DATA FOR FRAME DEFINTION AND MAINTENANCE

2.1 Current System and Proposed Systems

The current system for definition and maintenance of business survey frames is illustrated in figure 1A. It is cumbersome and somewhat fragmented. The principal component is the Business Register which provides the basic information from which separate frames are developed to meet the needs of individual surveys. The main source of updating information for the Business Register is the payroll deduction system at Revenue Canada. Data bases of corporate tax returns and of individual tax returns, also from Revenue Canada are independently maintained. Some surveys are based on frames developed from these sources, not the Business Register.

The problems with the current system may be summarized as follows. First, it is ineffective: there is undercoverage and duplication; and it is difficult to integrate the data processed by the separate surveys. Secondly, it is inefficient as a result of duplication of effort in the maintenance of individual survey frames.

In principle, the system illustrated in Figure 1A will be replaced by that in Figure 1B.

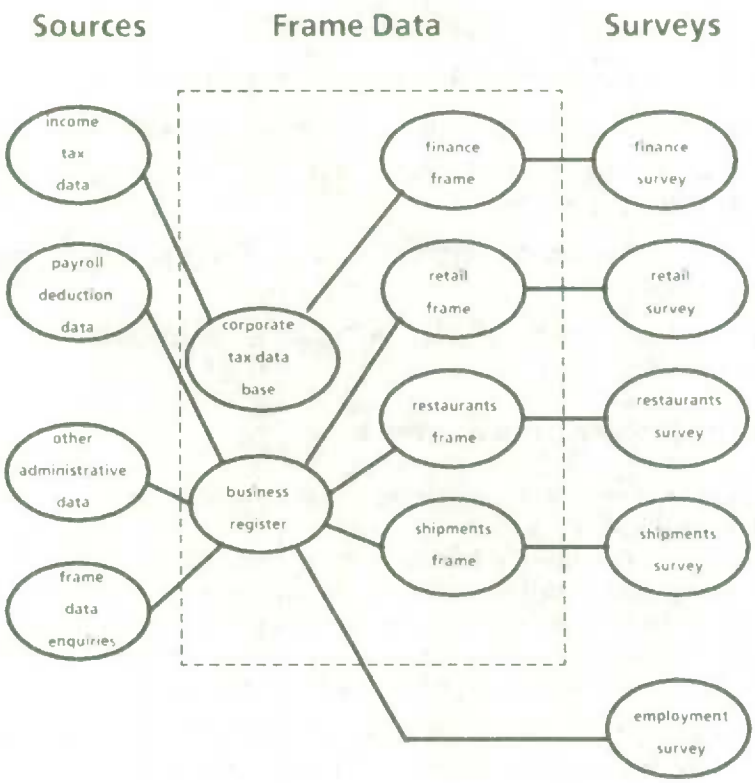


Figure 1A: Current

Figure 1: Systems for Provision of Frame Data (Simplified)

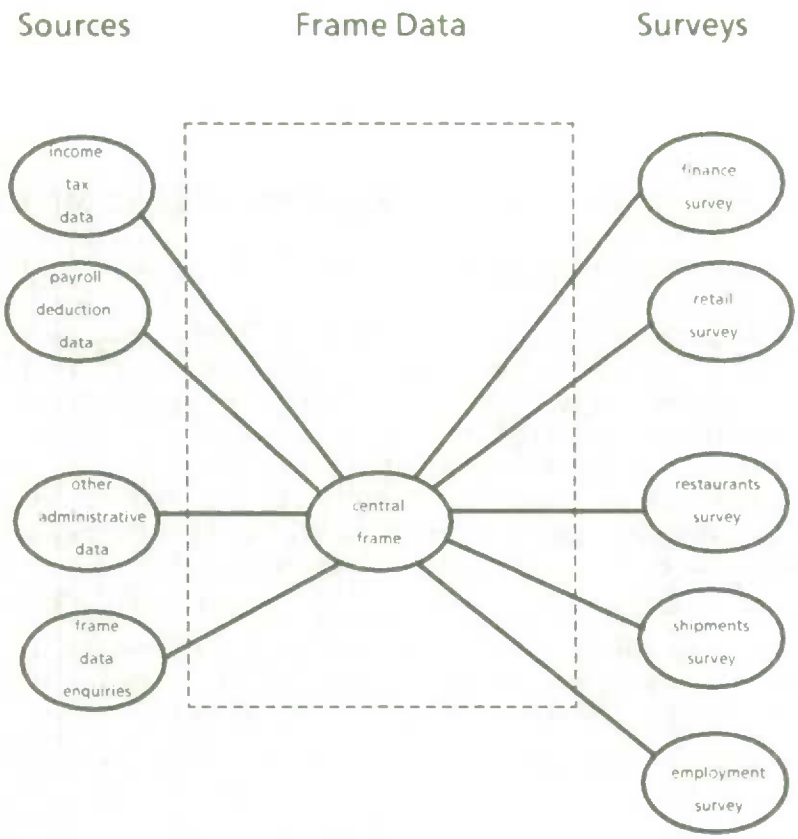


Figure 1B: Target

The main strategic elements in moving towards this solution are:

(a) development of standard concepts and an information model;

(b) implementation of a new central frame data storage system based on the standard concepts and information model;

(c) use of administrative data and the introduction of a "profiling" program, to create and maintain data in the base.

In the following subsection these elements are elaborated as a sequence of problems and corresponding solutions.

2.2 Frame Creation: Problems and Solutions

Problem. Survey data requirements determine the appropriate target unit. In other words, there is no single unit of which the target population for all business surveys can be composed. For example, consider a business with 3 retail sales outlets in two provinces, 2 branch offices handling the retail outlets, a wholesale branch office with 2 outlets, and a head office. Data on sales may be available for each of the individual outlets, whereas profit and loss statements may be obtainable only at branch level, and a consolidated balance sheet only from the head office for the whole business.

Solution. The approach adopted is to recognize explicitly the need for different types of statistical unit. A four level hierarchy of target statistical units is defined, thereby providing an appropriate unit for each type of survey. This is illustrated in Figure 2 which shows one "statistical enterprise" which can report consolidated data. The enterprise embraces two "statistical companies", from which unconsolidated financial statements can be obtained, three "statistical establishments" providing data leading to value added, and five "statistical locations" providing sales data.

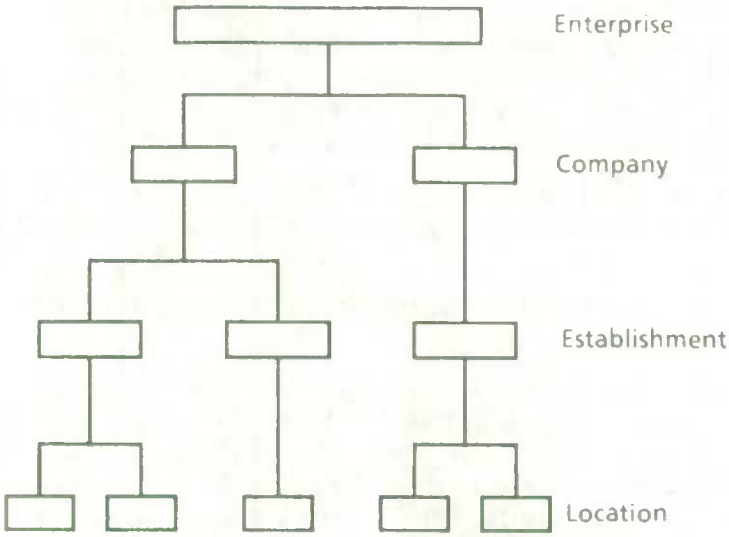


Figure 2: Statistical Representation of a Business in Terms of the Four Level Hierarchy

Problem. No lists of statistical units for use in business surveys exist a priori. They have to be created and maintained by Statistics Canada.

Solution. The approach adopted is to construct lists of statistical units based on administrative data. This is not a new approach. In fact, with one exception, every business survey frame has its origin in an administrative list, thereby demonstrating the fundamental dependence of business surveys upon administrative sources. The exception is an area frame developed by Statistics Canada and presently used to complement frame coverage for the Monthly Retail Trade Survey. Even this last vestige of a business frame which is not derived from an administrative source will disappear, at least on a trial basis, when the newly redesigned Retail Trade Survey is introduced.

Problem. Administrative and statistical units cannot be expected to coincide as they serve different needs. In particular, there are four levels of target statistical unit, according to the survey data being collected, at most one of which could be matched to the units provided by a given administrative source. Thus, in general, administrative lists cannot be used directly as survey frames.

Solution. The first step is to define an information model (Statistics Canada 1985) which explicitly recognizes the various types of unit in the business world and which relates these units to the statistical targets. The model incorporates five distinct unit types:

(a) legal - for example, incorporations under federal or provincial charter;

(b) adnministrative - for example, payroll deduction account holders, income tax filers;

(c) operating - for example, divisions, profit centres plants, etc., corresponding to the way in which the business organizes itself and keeps its operating accounts; the legal, administrative and operating units jointly define the view the business has of itself;

(d) statistical - the target units for statistical measurement purposes, i.e. the statistical agency's view of the business;

(e) reporting - providing the linkage between the statistical target units and the business operating units.

The second step is to define procedures whereby the sets of statistical units are derived from administrative lists. The procedures, referred to collectively as "profiling",

are as follows. The boundaries of a business are defined based on administrative data, eg., lists of tax filing corporations, and on ownership and control information from the Corporations and Labour Unions Reporting Act program. The operating structure for the business is then obtained by interview or questionnaire. Finally, the statistical structure is generated automatically from the operating information.

The third step is to implement the information model as a data base, termed the Central Frame Data Base (CFDB), capable of storing and manipulating all the information obtained by profiling.

Problem. Profiling is an expensive operation. It would be far too costly to "profile" every business in the country.

Solution. Two important characteristics of busineses have lead to a solution. First, a relatively small number of large businesses lend to dominate the economy; secondly, for small businesses, the administrative and the statistical units (all four levels) often coincide. Thus, the approach adopted is to profile large businesses in order to determine the corresponding statistical units, but for small businesses to use the administrative lists directly to provide frames. The CFDB is, therefore, divided into distinct parts, as shown in Figure 3A. The "Integrated Portion" (IP) contains a unique unduplicated list of the statistical units covering all large businesses together with the associated "structural" information, i.e. administrative, legal, operating and reporting data. It is derived by profiling, and it involves complete linkage and unduplication of all administrative and other input sources. The statistical units, generated from the operating structure, are fully classified in terms of industry, geography and size. The "Non-Integrated Portion" (NIP) contains the small units required to complement the frame. It is derived directly from administrative data assuming a one to one correspondence between statistical and administrative units. Very small units are defined as Out-of-Scope for survey purposes.

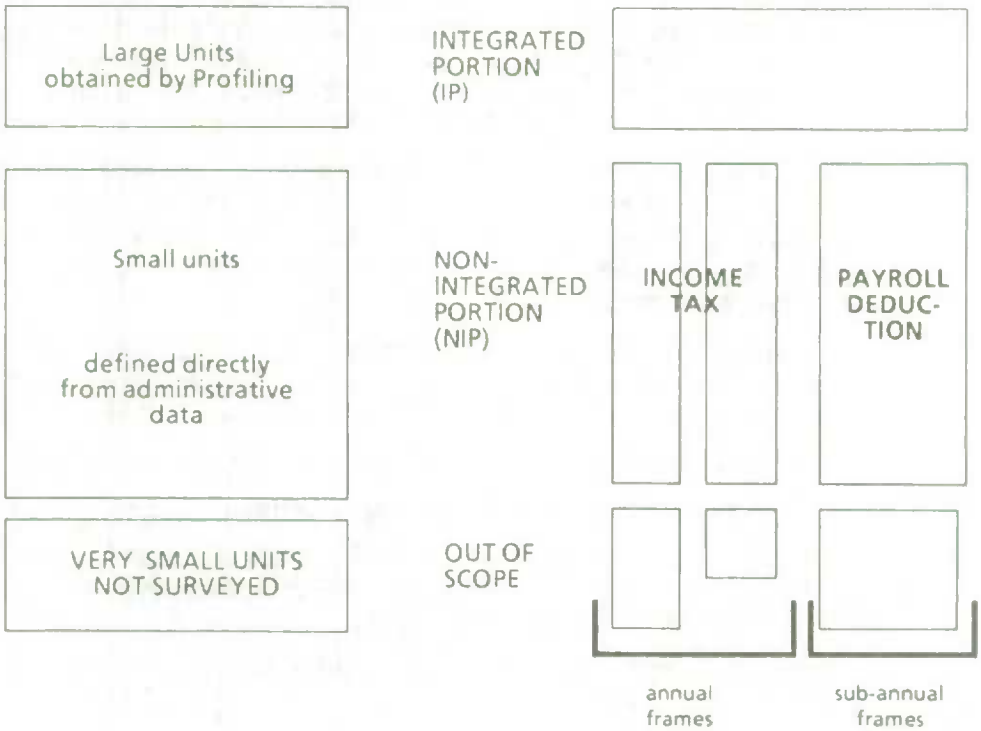


Figure 3A: Basic Concept

Figure 3B: Alternative Frames

Figure 3: Central Frame Data Base

Problem. There is no single administrative source which can produce a full complement of small units for the NIP. Income tax data provide a good basis for generating frames for annual production and financial surveys but involve too much time lag to be suitable for subannual surveys. The payroll deduction source provides more timely data, as appropriate for subannual surveys, but does not cover non-employers.

Solution. The approach is to use both these sources in constructing the NIP.

Problem. Many businesses are represented on both the payroll deduction and the income tax lists, thus there is the potential for considerable duplication if the lists are used together.

Solution. One approach would be to apply multiple frame techniques. This would add substantially to the complexity of estimation which would be undesirable in view of the many surveys involved.

A second approach would be to match and unduplicate the two lists. However, although both lists are provided by Revenue Canada, they do not share a common identification scheme, they are not presently linked and they cannot be readily matched and unduplicated. Automated linkage procedures based on name and address would tend to miss a considerable number of matches as payroll deduction name and address information can often differ significantly from income tax information for the same unit. The procedure would also generate a large number of potential matches requiring costly manual resolution.

The approach adopted is not to attempt to link the lists, i.e., to leave them non-integrated (hence the name), and to use one or the other to complement to the IP in forming a frame. This is illustrated in Figure 3B.

Problem. Neither of the two administrative sources provide sufficient information for the precise classification of units by industry, geography and size required for stratification and efficient sample selection. For example, it is desirable to stratify and sample units by 4 digit 1980 Standard Industrial Classification (Statistics Canada, 1980) for annual surveys of economic production. However, the data contained on a corporate tax return are sufficient to determine a 4 digit classification for only 74% of all returns. The corresponding figure for unincorporated tax returns indicating business income is 50%.

Solution. In the past, uncertainty about the industrial classification of tax returns was resolved by assignment of the most likely value. The errors introduced by this procedure have generated doubts about the quality and utility of tax returns. Thus the approach adopted for both payroll deduction and income tax data is always to contact a unit in case of uncertainty. In addition, to reduce costs further, a two-phase sample design will be employed for all surveys using the income tax based NIP so that precise industrial codes need be maintained only for the first phase sample (Colledge, Estevao and Foy, 1987).

Problem. When a sample is selected from the NIP it may contain some administrative units which are not in one-to-one correspondence with the appropriate statistical units, i.e., those which would have defined had the unit been profiled. For example, a single administrative unit might correspond to three businesses - an unincorporated tax return referring to a doctor who has professional income, is also a partner in an athletic club and owns a hobby farm. In this case, ideally, there would be three statistical units, not one. The converse situation can occur. A sampled tax unit may represent a tax filer in partnership with four others, i.e., ideally one statistical unit, but five tax units. For samples drawn from the payroll deduction based NIP, a similar lack of correspondence between the administrative and appropriate statistical units can exist.

Solution. The approach adopted for sampled tax units with multiple businesses is to define a separate statistical unit for each individual business, as identified by a separate set of financial statements attached to the tax returns. The sampling weight of the administrative unit is carried over to each such statistical unit.

Tax filer partnerships imply duplication on the tax frame. This is handled by multiplying the weight of each sampled unit by the share that unit has in the partnership. The same approach is adopted in the use of payroll deduction accounts referring to a single business.

Problem. Given that the CFDB has been created it must be updated to reflect, as far as possible, all relevant changes of structure and of classification which occur in the business world. This is a complex process. Updating information is available from a variety of different sources. The data obtained from any one source may be incomplete and may even be in partial conflict with information from other sources. There is a bewildering number of alternative ways in which the sets of legal, administrative, operating, and statistical units and their relationships can be updated, and there are some pitfalls to avoid. For example, changes of legal structure such as mergers, amalgamations, takeovers, creations of subsidiaries, etc., do not necessarily imply any changes in the corresponding operating or statistical structures. Thus, if the sets of statistical units were to be updated automatically as a result of information from administrative or legal sources, there might be a speciously high incidence of apparent "births" and "deaths" of statistical units, and an attendant risk of incomplete or duplicate coverage.

Solution. To handle this situation, a comprehensive set of fifty or so "standard events" which can take place in the business world have been defined (Armstrong et al, 1986). Any indication of a change is considered as a "signal" in response to which the corresponding units are investigated, and reprofiled if necessary, to deduce which, if any, of the fifty standard events have occured. CFDB updating is always in terms of these events. Precise definitions of births, deaths and changes to statistical entities are embedded in the rules for definition of standard events and for statistical entity generation. There are quite large numbers of signals, obtained from essentially three types of source: survey feedback; administrative processes; and CFDB routine reprofiling. These signals are placed on the CFDB "workbench" where they are sorted according to complexity, batched into work units, and automatically allocated to CFDB maintenance staff when they request work assignments. (See Figure 4).

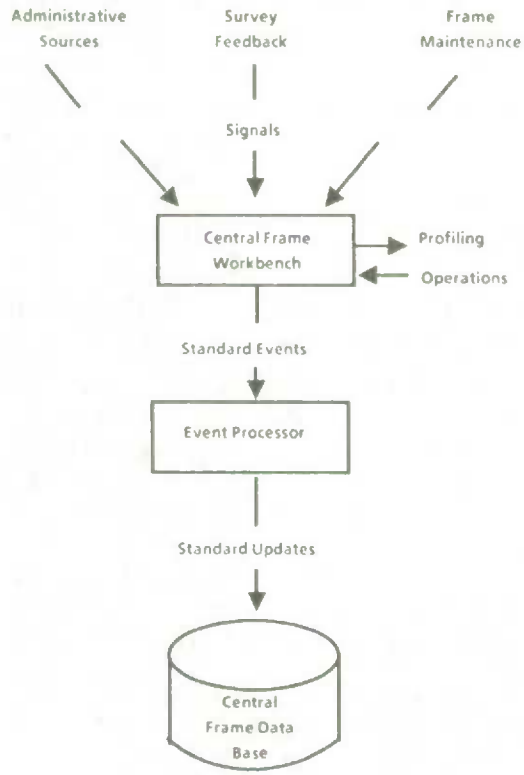


Figure 4: Central Frame Data Base Updating

In accordance with the general philosophy of different treatments for large and small businesses, the maintenance of NIP units consists of automated updating from the administrative files, supplemented by survey feedback and, where necessary, by direct contact.

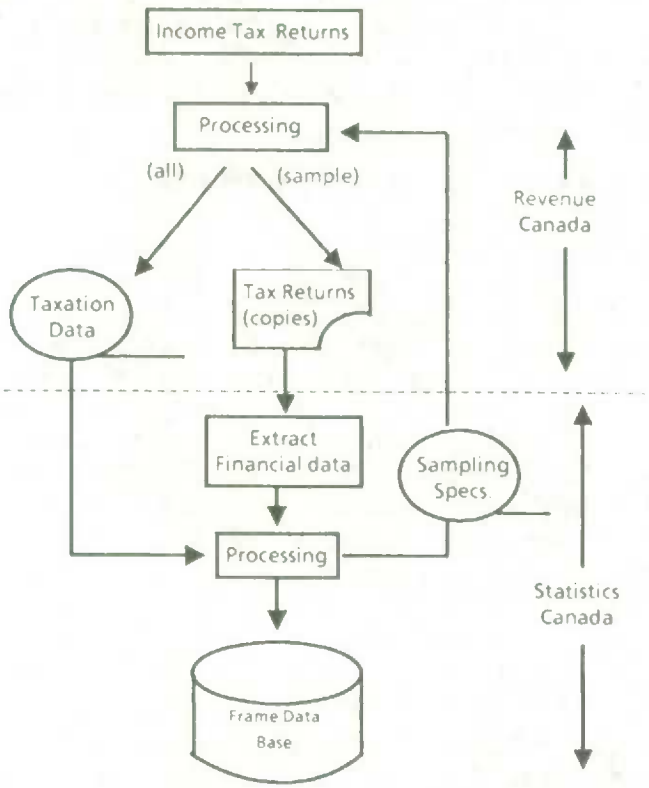
3. ADMINISTRATIVE DATA TO SUPPLEMENT SURVEY DATA

A principal objective of the Business Survey Redesign Project is to reduce operating costs and respondent burden through the use of administrative data. In this context the single, most significant application is the partial replacement of annual survey data by income tax data. Thus, the focal elements of the strategy are:

(a) the use of financial data from tax returns in place of survey data collection for small businesses;

(b) the coordination of income tax data sampling, acquisition and processing procedures to meet the collective needs of all surveys.

The general procedures by means of which income tax data are acquired from Revenue Canada, are illustrated in figure 5. Revenue Canada captures all taxation data items and certain financial items for every tax return. These data are made available in machine readable form. However, they do not have all the information required for business surveys. Thus, Statistics Canada requests a sample of tax returns, as identified by individual unit numbers or by sampling algorithm based on classes of unit. The specified returns are identified during processing at Revenue Canada, copied, and forwarded to Statistics Canada. The additional data items are then extracted from the schedules and financial statements attached to the returns. The use of these data by annual surveys of economic production and the annual survey of financial and taxation statistics are described in the following subsections.



3.1 Annual Survey of Economic Production

Currently, annual production data are collected by a number of independent survey operations, covering separate industry groups. Examples of these operations are the Annual Survey of Manufactures, the Census of Construction, the Wholesale Trade Survey, the Retail Trade Survey, the Annual Services Program, etc. The strategic objective of the Business Survey Redesign Project is to standardize the concepts and to coordinate procedures over all these surveys so they can be collectively viewed as a single "Annual Survey of Economic Production". Some specific standard concepts are described in the following paragraphs.

The target statistical unit is the establishment. A set of data items at least sufficient for the computation of "census value added" (Statistics Canada, 1980) is collected at the finest level of geographic and industrial detail that can be supported on an annual basis.

The frame is derived from the IP and the income tax based NIP. All units with fiscal year ending between April 1, Y and March 31, Y+1 are defined to be in-scope for reference year Y.

Data are obtained from the IP units by direct contact using a personalized questionnaire. To reduce respondent burden and reporting error, data are requested for the respondent's fiscal year (rather than the calendar year). Where statistical and tax units coincide, income tax returns may be used as the source of financial data in place of direct collection.

For the NIP units all financial items are obtained from income tax returns. These data are supplemented by direct survey of other characteristics such as employment, commodities produced or consumed, types of service provided, etc. As income tax units on the frame cannot be reliably classified to 4 digit 1980 SIC a two phase sample design is used, as previously noted.

Year to year sample overlap is made as large as possible to provide a good basis for year to year comparisons. In this context, respondent burden is not a major consideration as most of the data are derived from tax returns. Sample overlap is controlled by selection using a reproducible random number obtained by "hash" function from the tax unit identifier (see Sunter, 1986).

3.2 Annual Survey of Financial and Taxation Statistics

Currently, the survey is based exclusively upon income tax data. Financial items for the universe of corporations available from Revenue Canada in machine readable form are merged with more detailed data extracted from a sample of tax returns. The problems with this approach are threefold. First, data for corporations which are under common ownership and central are not consolidated, and are thus difficult to relate to financial information obtained by the quarterly financial survey from consolidated units. Secondly, the data are not easy to match at micro level to data from economic production surveys as the corresponding sets of units do not have a well defined relationship to one another. Thirdly, the sample of income tax returns is not coordinated with the sample used for the economic production surveys, implying duplication of effort and loss of information.

The elements of the strategy are:

(a) to define the target unit to be the statistical enterprise as defined for the CFDB, not the corporate tax unit, and to derive the frame from the CFDB rather than the Revenue Canada corporations file;

(b) for large units to obtain the required financial data using four quarters of information from the quarterly survey, i.e. to move away from the use of administrative data, in order to obtain consolidated information;

(c) to use a sample design similar to that for the annual survey of economic production, including the two phase approach for NIP units.

4. OTHER USES OF ADMINISTRATIVE DATA

In this section other uses of administative data in the context of business survey redesign are described.

The inadequacies of the present business survey systems and procedures at Statistics Canada stem primarily from the fact that there are many survey operations which have developed essentially independently of one another. The data outputs are difficult to integrate, and the maintenance or redesign of the tailor-made systems is expensive. The elements of the BSRP strategy to address these problems may be summarized briefly as follows.

First, there is a review and rationalization of objectives for the whole program of business surveys. In general, subannual surveys are to focus on estimates of change, and on timeliness, rather than on detail which is to be provided by annual surveys. Secondly, there is a standardization of concepts and development of generic methods and generalized systems which will relieve respondent burden, facilitate data integration and reduce redesign and maintenance costs. In particular the existing program of surveys will be realigned to the new procedures for the provision of frame data and acquisition and use of income tax data outlined in the previous sections.

In this context, administrative data are of considerable importance, but from a survey manager's perspective, their role is essentially transparent. For example, although administrative sources are a key element in the construction of maintenance of frame data, each survey manager simply receives a frame from the CFDB, and has no occasion to interact directly with, or to query the (administrative) sources of that frame. Likewise the survey manager receives income tax data, at unit or aggregate level, from the central tax data processing function in the same way as data from survey questionnaires may be obtained from a central data capture function. Again there is no direct interaction with the administrative source.

4.1 Generic Function: Use of Administrative Data

The generic functions and data bases associated with a business survey may be summarized as in Figure 6. The breakdown of the entire survey process into these ten generic functions is somewhat arbitrary but convenient for description purposes. Colledge and Lussier (1987) provide more details. In the following paragraphs, the role played by administrative data is outlined, taking the functions in sequence.

The use of administrative data for frame creation and maintenance has been extensively discussed (Section 2). For sample stratification and allocation, administrative data may be used as a proxy when no relevant survey data are available, eg., in designing a completely new survey. Sample allocation and selection are carried out essentially independently of whether data collection is to be by survey or from income tax return. The generation of reporting units and contact materials (questionnaires, control lists, etc.) follow the same principles whether the data are to be obtained by survey questionnaire or from income tax, except that, in place of questionnaires, lists of tax units are sent to Revenue Canada for interception of the corresponding tax returns. Similarly, data capture, editing and follow-up procedures are in accordance with the same principles regardless of data source although follow-up requests for income tax data go to Revenue Canada rather than to businesses.

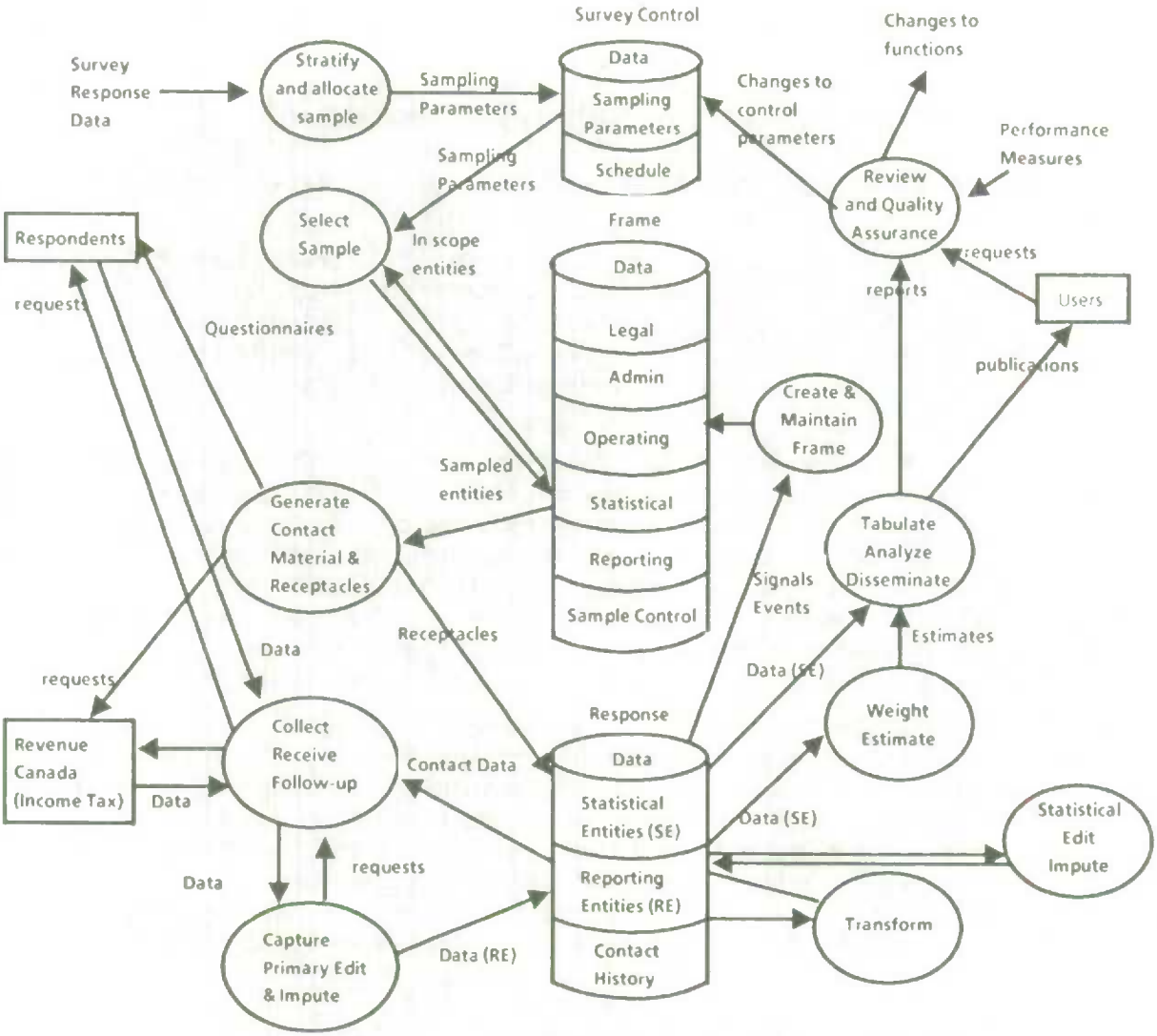


Figure 6: Generic Functions, Data Flows and Storage (Simplified, and Excluding Frame Creation and Maintenance Details)

The "transformation" procedure is based on a relatively new concept. Its function is to convert data **from** reporting units to data **for** the target statistical units, thereby facilitating the interpretation of data across surveys. This is in contrast to existing procedures where survey specific reporting unit data are the final product. In the case of income tax data this transformation is handled at the time the statistical units associated with the first phase tax sample are being identified from the selected tax returns.

Edit and imputation includes the use of administrative data from machine readable files or from the tax sample to impute for missing survey data. Administrative data may also play a role as auxiliary data at the sample weighting and estimation stage, for example, by providing benchmark totals for the Labour Income Program. The use of payroll deduction remittance data for the monthly Survey of Employment, Payroll and Hours (Cotton, 1987) is also being considered.

Administrative data are valuable for analysis, evaluation and audit purposes. As regards tabulation and dissemination, any constraints imposed by administrative agencies on the use of their data must be taken into account. This may restrict the dissemination

which would otherwise be possible. For example survey micro data can be released to the provincial governments, but unit level income tax data cannot.

In terms of quality assurance, administrative procedures cannot be subject to the same degree of quality control by Statistics Canada as can data processing within the agency itself. Thus the quality of incoming administrative data has to be carefully measured and the corresponding source advised immediately of any shortfalls. The administrative procedures themselves have to be monitored to ensure that changes are not made without full knowledge of the effect they will have upon the statistical program. Ongoing liaison with the administrative agencies is vital (Gates, 1987).

5. CONCLUSION

The paper contains a review of the ongoing Business Survey Redesign Project at Statistics Canada from the perspective of using administrative data. The applications of administrative records have been categorized and described under three headings: frame definition and maintenance; partial replacement of annual survey data collection by income tax data; and other uses in the survey process. Particular reference has been made to the two principal sources, namely payroll deduction and income tax. The main conclusions are summarized below.

In terms of defining the Central Frame Data Base (CFDB) to serve the whole program of business surveys, administrative data play a vital role. They provide the basis for all survey frames. Without them frame construction costs would be astronomical. Administrative units are used directly as statistical units for sampling small businesses, but for large businesses this approach is not appropriate. Thus there are differential procedures according to unit size. For large businesses, administrative data are the starting point for the process ("profiling") of delineating statistical units. For the complement of small businesses no single administrative source gives suitable coverage, so the Non-Integrated Portion (NIP) of the CFDB contains alternative frames based on two separate administrative sources, payroll deduction and income tax. The classification information required for stratification of the NIP is not available for all administrative units. It has to be obtained by a supplementary process involving direct contact of the units. As this is expensive, a two-phase design is used for sampling the income tax based NIP so that only units selected in the first phase sample need to be fully classified. The distinctions which, ideally, should exist between administrative and statistical NIP units are recoginized only for units selected in the first phase tax sample or other survey sample.

CFDB maintenance is based on processing administrative records and survey feedback together with information from specific frame reveiw procedures. Indications of change, "signals", are identified and interpreted as one or more "standard events", in terms of which the CFDB is updated. For NIP units the conversion of administrative signals into standard events is automated, but for IP units clerical investigation is required. Automation of the procedures for processing administrative signals has been considered but without success to date. It would produce substantial savings in clerical resources.

In the context of replacing direct survey data collection by administrative data and hence reducing repondent burden and operating costs, the most notable development is the use of income tax data. Under the BSRP strategy, income tax returns are the source of financial data items for all small units in-scope for the annual survey of economic production. Direct contact of selected small units occurs, but is restricted to collection of additional information regarding industrial classification (when the income tax data are insufficient), and other, non-financial characteristics. For the annual survey of corporate financial and taxation statistics, income tax returns are the sole source of all data, except financial items for the very largest corporations. In the case of these corporations the

administrative unit is not appropriate and information from the quarterly financial survey are used. This is the only instance in recent years where the planned use of administrative data has actually been reduced.

It is difficult to evaluate the quality of income tax data as a replacement for survey data. Studies comparing data values derived from income tax returns and from annual survey questionnaires indicate substantial discrepancies between the two sources but it is not easy to determine which source provides the more "accurate" values due to the sensitivity of respondents to enquiries regarding their income tax returns. The view has been taken that the values appearing on the financial statements prepared for income tax returns are more likely to be accurate than survey responses, but that the tax data will not always provide the precisely breakdown of financial items required.

In addition to these major applications, administrative data are used to impute for missing survey data, to provide benchmark totals for sample allocation and/or estimation and to assist in survey evaluation and audit. In summary, administrative data play a very important role in the business survey program, a role which can be expected to continue expanding.

REFERENCES

Armstrong, G., Monty, A., Woods (1986). "Definitions of Standard Events", Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

Bankier, M., Bordeleau, C., Carruthers, I., Demmons, P., Finlay, M., and Leduc, J., (1985). "Preliminary Report on Documentation of Administrative Data Sources", Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa

Beckstead, D., and al (1985). "PD Processing Time Lag Study", Business Register Working Paper, Statistics Canada, Ottawa.

Brackstone, G.J. (1987). Issues in the Use of Administrative Records for Statistical Purposes, Presented at the Statistical Society of Canada Annual Meeting, 1987, Québec City.

Cain J., and al (1984). "Infrastructure Development, Objectives, Policy and Strategy", Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

Cochran, W.G. (1977). "Sampling Techniques", Wiley, New York.

Clark, C., and Lussier, R. (1987). "The Use of Administrative Data for Initial and Subsequent Profiles of Economic Entities", *Proceedings of the Section on Survey Methods, 1987*, American Statistical Association, Washington.

Colledge, M. (1987). "The Business Survey Redesign Project - Implementation of a New Strategy of Statistics Canada", presented at the Bureau of the Census Third Annual Research Conference, Washington.

Colledge, M., and Lussier, R. (1985). "A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys", *Proceedings of the Section on Survey Methods, 1985*, American Statistical Association, Washington.

Colledge, M., Estevao, V., and Foy, P. (1987). "Experiences in Coding and Sampling Administrative Data", Proceedings of the section on Survey Methods 1987 (to be published). American Statistical Association, Washington.

Estevao, V., Ambroise, P., and Colledge, M. (1983) "A study on the Quality of Certain Fields of the Business Register Master File", Business Register Working Paper, July 1983, Statistics Canada, Ottawa.

Estevao, V., and Tremblay, J. (1985). "A Report on the Quality of the Data in the BRMF - SARUS Study" - 1984/85.

Estevao, V., and Tremblay, J. (1986a). "An Evaluation of the Assignment of Standard Industrial Codes from PD-20 Data", Business Survey Redesign Project Working Paper, May 1986, Statistics Canada Ottawa.

Estevao, V., and Surman, P. (1987b). "A Study on the Use of Research Information to Obtain Complete SIC Codes for Incorporated Businesses", Business Survey Redesign Project Working Paper, March 1987. Statistics Canada, Ottawa.

Estevao, V. and Tremblay J. (1986b). "An Evaluation of the Assignment of Standard Industrial Codes from T2 Tax Data", Business Survey Redesign Project Working Paper, November 1986, Statistics Canada, Ottawa.

Foy, P. (1987). "Two Phase Sample Design for Estimation from Tax Data for Annual Surveys of Economic Production", Business Survey Redesign Working Paper, September 1987, Statistics Canada, Ottawa.

Foy, P., and Corriveau, P. (1986). "Evaluation preliminaire de l'emploi d'un échantillon maître des comptes PD dans la partie non-intégrée du CFDB" Business Survey Redesign Project Working Paper, March 1986, Statistics Canada, Ottawa.

Foy, P. (1987). "Development of the OC Capability for the Annual Surveys of Economic Production", July 1987, Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

Hostetter, S.C. (1983). "The Verification Method on a Solution to the Industry Coding Problem", *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Washington.

Konschnik, C., Monsour, N., and Detlefsen, R. (1985). "Constructing and Maintaining Frames and Samples for Business Surveys". *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington.

Lussier, R., and Colledge, M. (1987). "Business Survey Redesign Project Quality Assurance Strategy", Business Survey Redesign Project Working Paper, July 1987, Statistics Canada, Ottawa.

Statistics Canada, (1970). "Standard Industrial Classification 1970, Catalogue 12-501E, Statistics Canada. Ottawa.

Statistics Canada, (1980). "Standard Industrial Classification 1980", Catalogue 12-501E, Statistics Canada, Ottawa.

Sunter, A. (1986). "Implicit Longitudinal Sampling from Administrative Files: A Useful Technique", *Journal of Official Statistics*, Vol. 2, No. 2, 161-168.

Sunter, A. (1987). (1987). "A Note on the Tax Universe Master Sample", Business Survey Redesign Project Working Paper, June 1987, Statistics Canada, Ottawa.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

SMALL AREA ESTIMATION OF EMPLOYMENT IN DIFFERENT CLASSES OF HOURS WORKED

SIXTEN LUNDSTRÖM[1]

ABSTRACT

Until 1980 the Swedish Population and Housing Census form has included questions about employment, but from the 1985 Census employment data are planned to be collected mainly via registers. One shortcoming of this new data source is that it does not provide any information about the number of hours worked. Therefore, we have tried to develop a small area estimator by combining data from the Labour Force Survey and a tax register. The paper describes the simulation studies of, among others, a synthetic and a logit model estimator and also the work on improving the estimator subsequently chosen.

1. INTRODUCTION

Many different users, in both the private and public sector, need annual information about the hours worked by economically active individuals. Municiple authorities have a great need for this kind of data when planning day care facilities, evaluating the need for public transport and forecasting employment. Previously, the Swedish Population and Housing Census provided such data every fifth year, but from the 1985 Census and on employment data will be collected via a system of registers. These registers, however do not provide information about hours worked. The domains are so small that conventional estimators based on nationwide samples do not give reliable estimates and therefore we have used auxiliary information to develop a model-dependent estimator or, as it sometimes is called, a synthetic estimator. In the future we will probably not have the opportunity to conduct thorough evaluation studies so it is important to develop an estimator that performs well over time. At Statistics Sweden model-based estimation is still relatively uncommon. To avoid confusion, we prefer the estimator to be as uncomplicated as possible.

2. DATA SOURCES

The data we would like to use when applying our estimate is taken from two sources namely, the Statistics on Regional Employment (SORE) and the Labour Force Survey (LFS). We might also use data from the 1980 census. SORE is based on a combination of several registers of which the most important is the tax register. Each year employers send income statements for each employee to the tax authorities. Statistics Sweden gets a copy of this register and merges it with the Register of Entreprises and the Register of the Total Population. With this we can then produce statistics about the economically active population. One shortcoming is that we do not have information about hours

[1] Sixten Lundström, Statistics Sweden, U/STM-0, 701 89, Örebro, Sweden.

worked. The LFS, on the other hand, collects information about hours worked. Each month the sample contains about 12,000 economically active persons. The LFS has variables that are also found in the SORE. For developing an estimator of hours worked, the most important variables are sex, age, income and industry.

The Census has measured the employment status for a given week in November every fifth year. The users would like to have an unbroken series of data even though a new data collection method is introduced. This is not easy to accomplish. Regarding the study variable "hours worked per week" we will present some results from a comparison between the census and the LFS. The SORE has a problem in isolating the economically active population but that problem is not dealt with here.

3. ESTIMATORS

The number of small area estimators is large and thus, cannot be compared in the same study. We have restricted this study to the estimators we are most confident about and for comparison purposes, a conventional estimator is included.

Suppose a population of size N consists of Q mutually exclusive and exhaustive small areas labelled $q = 1, \ldots, Q$. Moreover, suppose that each area can be further classified into H mutually exclusive and exhaustive classes by using a combination of some or all of the variables sex, age, income and industry, labelled $h = 1, \ldots, H$. The labelling gives a two-way cross-classification into HQ cells with N_{hq} population members in the hq-th cell, with a corresponding sample count n_{hq} in a simple random sample of size n. For aggregation across the subscript, we replace that subscript with '.'.

We want to estimate the percent of economically active persons in a given class of hours worked:

$$T_q = \frac{100}{N_{.q}} \sum_{i=1}^{N_{.q}} y_i, \tag{1}$$

where

$$y_i = \begin{cases} 1 \text{ if the i-th person belongs to the given class of hours worked} \\ 0 \text{ otherwise} \end{cases}$$

The synthetic estimator, SYNT, assumes that within the h-th subgroup the Q small area means are approximately equal. The expected percentage in area q is then given by

$$\text{SYNT} = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq} \hat{\bar{Y}}_{h.}, \tag{2}$$

where

$$\hat{\bar{Y}}_{h.} = \sum_{i=1}^{n_{h.}} y_i / n_{h.}$$

Unlike the SYNT-estimator, the following estimator uses totals of income instead of counts:

$$\text{SYNT RATIO} = \frac{100}{N_{.q}} \sum_{h=1}^{H} X_{hq} \sum_{i}^{n_{h.}} y_i / \sum_{i}^{n_{h.}} x_i, \tag{3}$$

where

x_i is the income for the i-th person

and

$$X_{hq} = \sum_{i=1}^{N_{hq}} x_i$$

It seems plausible to assume that as income increases, so does the probability of being employed more than a specified number of hours. It also seems plausible that this probability increases more at the middle of the income scale rather than at the bottom or top of the scale. Thus, a logit model is fairly obvious.

We estimated the following regression coefficients using weighted least squares.

$$u_h = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon ,$$

where

$$u_h = {}^{e}\log \frac{p_h}{1-p_h} ,$$

and

p_h = the probability that a person in subgroup h belongs to a specific class of hours worked.

The matrix $\underset{\sim}{x}' = (x_1, \ldots, x_m)$ denotes a set of dummy variables for categorical variables such as sex, age (in classes), industry, and a variable for the mean income in each class. When estimating the coefficients, the observed proportion belonging to a specific class of hours worked in subgroup h is used as an estimate of p_h.

The estimator, LOGIT, is given by:

$$\text{LOGIT} = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq}\, p_h^{\star} , \tag{4}$$

where $p_h^{\star}$ is the predicted proportion in subgroup h and is computed from

$$p_h^{\star} = e^{u_h^{\star}}/(1+e^{u_h^{\star}}),$$

and

$$u_h^{\star} = \hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_m x_m .$$

Remark: In the methodology study, Section 4, the LOGIT-estimator is used for different classes of hours worked. Throughout this study we consider hours worked a two-category problem — the given class and all the other classes.

Previously, Statistics Sweden has examined a model-dependent estimator for the number of economically active persons in small areas during the intercensal period. This estimator is called SPINK-estimator and is defined as follows.

We assume that the population is sorted by the income variable x. The estimator then has the following form.

$$SPINK = \frac{100}{N_{.q}} \sum_{h=1}^{H} \sum_{i=1}^{N_{hq}} z_{hi}, \tag{5}$$

where

$$z_{hi} = \begin{cases} 1 & \text{if the i-th person belongs to subgroup h and has an income in the interval } [\hat{G}_{1h}, \hat{G}_{2h}] \\ 0 & \text{otherwise} \end{cases}$$

G_{1h} and G_{2h} denote the lower and upper limits for what is a "reasonable" value for an individual's income in a given class of hours worked. The way these limits are estimated is explained in the following text and illustrated by Figure 1.

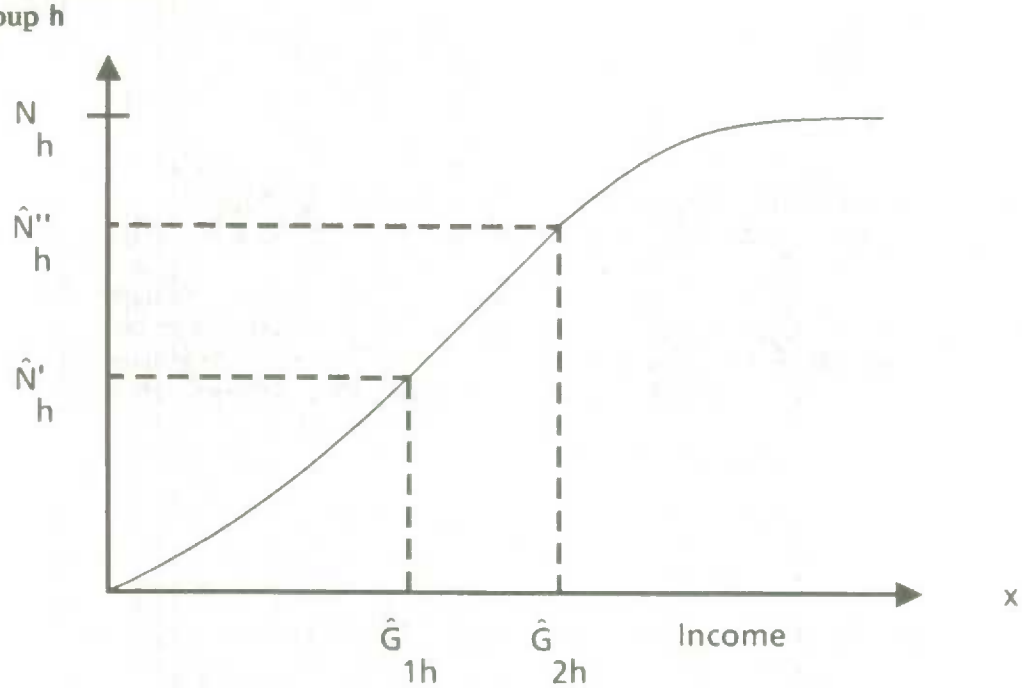


Figure 1: Number of individuals in subgroup h having an income less than x

$\hat{N}'_h$ denotes the estimated number of population members that belong to the classes of hours worked lower the given one. The number of persons belonging to the given class, N_{Gh}, is estimated by:

$$\hat{N}_{Gh} = N_{h.} \sum_{i=1}^{n_{h.}} y_i / n_{h.}$$

and

$$\hat{N}''_h = \hat{N}'_h + \hat{N}_{Gh}.$$

The income limits $\hat{G}_{1h}$ and $\hat{G}_{2h}$ are then estimated by projection on the income axis.

Särndal (1984) has, through the generalized regression method, developed an asymptotically unbiased estimator. Formally, it consists of two parts, the synthetic estimator (SYNT) minus an estimate of the synthetic estimator bias. The bias is often affected by a large sampling error. To minimize the mean square error, Cassel (1984) suggests (under specific assumptions) that the bias be multiplied by a constant, α.

The Optimal Corrected Synthetic Estimator (KORRSYNT) has the following form.

$$\text{KORRSYNT} = 100 \left(\sum_{h=1}^{H} N_{hq} \hat{\bar{Y}}_{h.} - \hat{\alpha} \frac{N}{n} \sum_{i=1}^{n_{.q}} e_i \right) / N_{.q}, \tag{6}$$

where

$$e_i = (y_i - \hat{\bar{Y}}_{h.}) \text{ for units belonging to subgroup } h,$$

and

$$\hat{\alpha} = 1 - \text{var}\,(\hat{A}_c)/\hat{A}_c^2,$$

where

$$\hat{A}_c = \frac{N}{nN_{.q}} \sum_{i=1}^{n_{.q}} e_i.$$

The previously mentioned estimators are model-dependent and consequently biased (the KORRSYNT-estimator is asymptotically design unbiased). We wanted to compare our model-dependent estimators with the following conventional unbiased direct estimator (DIR). It uses the sample mean in area q without using any supplementary information:

$$\text{DIR} = 100 \sum_{i=1}^{n_{.q}} y_i / n_{.q}. \tag{7}$$

4. DESCRIPTION OF THE SIMULATION STUDY

Any evaluation of model-dependent small area estimators is problematic. To estimate quantities like bias and mean square error, one needs true values. Usually this kind of estimator is so complicated that simulations have to be carried out. Due to scarce financial resources the number and types of small areas included in a study and the sample size are smaller than one should like them to be. This study suffers from these general difficulties and more specific difficulties that are linked to the surveys from which we took our data.

We expect a correlation between hours worked per week in November and annual income and hope to use the correlation as a bias-reducing factor in the estimation procedure. Some people work only part of the year and thus, will have small annual incomes. The tax register contains information about the period worked but, unfortunately, the quality of the data is poor. In the future, when the data collection procedure is improved, we will have better data.

4.1 Error Measurements

The mean square error (MSE) contains the sampling variability and the bias of the estimator and has the following form.

$$MSE(\hat{T}_q) = Var(\hat{T}_q) + B^2(\hat{T}_q), \tag{8}$$

where

$Var(\hat{T}_q)$ is the variance and $B(\hat{T}_q)$ is the bias of the estimator $\hat{T}_q$.

As part of this study we estimated $RMSE(\hat{T}_q) = (MSE(\hat{T}_q)^{\frac{1}{2}}$, $Var(\hat{T}_q)$ and $B(\hat{T}_q)$. Moreover, we estimated rel - |B|, which is given by

$$rel - |B| = 100 \cdot |B(\hat{T}_q)|/RMSE(\hat{T}_q). \tag{9}$$

To have a summary measure we also calculated the means (over small areas) of the error components. To compute these values for the various estimators discussed in the previous section, a Monte Carlo simulation study was carried out. For each sample r (r=1, ..., R) we calculated the estimator value $\hat{T}_{q(r)}$ and finally

$$B(\hat{T}_q) = (\sum_{r=1}^{R} \hat{T}_{q(r)}/R) - T_q, \tag{10}$$

and

$$Var(\hat{T}_q) = \left[\sum_{r=1}^{R} \hat{T}^2_{q(r)} - (\sum_{r=1}^{R} \hat{T}_{q(r)})^2/R \right] / R. \tag{11}$$

From the formulas above we could calculate all of the error components.

4.2 Sample Sizes

The model-dependent estimators use sample information for each subgroup h to estimate, the number of economically active persons in a class of hours worked. The sampling variability is then expected to be rather small even if the sample size is small, say 1,000. We can expect the bias to be a large part of the total error.

The design-unbiased estimator DIR is based only on the observations in the small area and is much more dependent on the sample size than the model-dependent estimators are.

Two different sample sizes, 1,000 and 2,500, are used in the simulation study. Both sample sizes will give a little too large sampling error for the model-dependent estimators and too small sampling error for DIR compared to what would be obtained when using regular LFS data.

4.3 Results

The first minipopulation in the study consisted of the LFS sample for October, November and December 1980 and the out-of-sample panels for all of the preceeding of

that year. We tried to exclude those not working the whole year by using registers, but this screening was not entirely successful. We also tried to exclude the employers because this group will probably not belong to the SORE population.

We merged that population with the 1980 census register and thus received information about hours worked both from the LFS and the census.

After these reductions the study population consisted of about 35,000 economically active persons. This population is too small to carry out a study of estimates on the municipality level, so we decided to use counties instead. The register of the population contained information about sex, income, hours worked (from the LFS and the Census) and a county code.

We have carried out simulations for different classes of hours worked, for different sets of income classes, and for two sample sizes. The analysis of the results leads to the same conclusion and for this reason we present only one of the simulation alternatives.

In Table 1 we present the simulation estimates of RMSE when sample size is 1,000 and when H=16 (sex * income classes). The parameter T_q, is the percent economically active persons working more than 34 hours per week. The different counties have a T_q-value in the interval 68-80.

Table 1

Estimated RMSE $(\hat{T}_q)$ for different estimators, when n = 1,000. $(68<T_q<80)$

COUNTY	SYNT	SYNTRATIO	LOGIT	SPINK	KORRSYNT	DIR
01	1.59	4.68	1.42	3.36	1.88	3.29
03	2.18	2.98	1.98	1.99	3.69	8.13
04	1.71	1.34	1.42	1.95	3.46	7.85
05	1.47	1.30	1.34	1.45	3.28	7.43
06	1.32	1.37	1.56	1.88	2.72	7.46
07	1.32	1.63	1.44	2.06	3.07	7.59
08	1.42	1.32	1.36	1.72	3.40	8.51
09	3.82	4.19	4.51	5.84	4.85	8.13
10	1.29	2.04	2.00	1.41	3.37	8.03
11	1.36	1.71	1.75	2.66	2.98	8.30
12	1.33	1.56	1.31	1.39	2.44	5.26
13	1.33	1.53	1.27	1.54	3.19	8.51
14	1.58	2.36	1.32	2.31	2.63	5.67
15	2.41	2.44	3.01	2.79	3.39	7.15
16	2.03	2.64	2.76	2.41	3.88	8.29
17	1.71	2.74	1.90	3.00	3.32	7.72
18	1.63	1.84	2.26	1.68	2.79	7.61
19	1.30	1.31	1.49	1.52	2.99	6.82
20	2.15	2.76	2.90	2.73	3.93	7.86
21	1.28	1.39	1.80	1.42	3.28	7.68
22	1.66	2.21	1.93	1.60	3.00	7.90
23	2.29	2.57	3.25	2.70	3.52	7.97
24	1.49	1.40	1.30	1.45	3.67	7.87
25	4.76	3.71	3.60	5.60	5.02	8.52
Mean	1.85	2.21	2.04	2.35	3.32	7.48

Table 1 indicates that SYNT is the best estimator when using the error measurement RMSE. That conclusion will also follow from other simulations carried out on this minipopulation. The unbiased estimator, DIR, and the asymptotically unbiased estimator, KORRSYNT, suffer from a large RMSE, which is mainly explained by the sampling variability. Indirectly, that will be shown in Table 2 where we present the estimated bias for the different estimators.

Table 2

Estimated $B(\hat{T}_q)$ for different estimators

COUNTY	SYNT	SYNTRATIO	LOGIT	SPINK	KORRSYNT	DIR
01	.83	4.45	-.28	3.10	.55	.09
03	1.78	2.65	1.50	1.57	1.04	-.17
04	1.17	.45	.52	1.39	.72	-.77
05	.76	-.03	.27	.17	.51	.92
06	-.41	-.49	-.87	-1.32	-.24	-.39
07	-.45	-.98	-.66	-1.57	.04	.63
08	.65	.02	.26	-1.23	.27	-.68
09	-3.58	-3.95	-4.30	-5.66	-1.71	.52
10	-.25	-1.54	-1.49	-.47	-.06	-.62
11	-.52	-1.06	-1.18	-2.34	-.37	-.83
12	.44	.73	-.23	.01	.34	.15
13	.45	.80	-.01	-.62	.76	.45
14	.92	1.93	.17	-1.95	.66	.30
15	-2.05	-2.07	-2.72	-2.32	-.82	-.27
16	-1.58	-2.30	-2.43	-2.05	-.59	.74
17	-1.15	-2.40	-1.37	-2.63	-.26	.52
18	-1.04	-1.32	-1.85	-1.11	-.59	.40
19	.25	.12	-.69	.11	.45	.86
20	-1.73	-2.42	-2.57	-2.31	-.60	.37
21	-.21	-.45	-1.24	-.59	.04	-.30
22	-1.08	-1.81	-1.41	-.91	-.43	.66
23	-1.87	-2.19	-2.96	-2.20	-1.05	-.60
24	.79	-.57	.08	-.17	.64	-.34
25	4.57	3.46	3.34	5.41	2.27	-.88
Mean of absolute values	1.19	1.59	1.35	1.72	.63	.52

Table 2 shows that both the size and the sign of the bias of the model-dependent estimators vary strongly between different counties. The counties with the largest biases are Gotland (09) and Norrbotten (25). Gotland is an island with a large proportion of farmers (low assessed income but high numbers of hours worked). Norrbotten is the northernmost country in Sweden with many miners (relatively high assessed income). When we included the industry variable (see Section 4.4) we noticed that the biases were reduced.

The bias is the dominating error of the model-dependent estimators even in the case of only 1,000 observations in the sample. That can be found from a comparison of Table 1 and Table 2. For the SYNT-estimator the mean of rel - |B| is 56 percent.

In one case we have used a sample size of 2,500 individuals and then observed that, as expected, the RMSE of estimators KORRSYNT and DIR decrease more than the model-dependent estimators comparing the results of the 1,000 and 2,500 sample size.

As mentioned before, one problem with combining different data sources is that we will have different definitions of the study variable, different data collection methods, etc. In this minipopulation we are able to compare the values of the study variable, i.e., hours worked per week, from the LFS and the census. In Table 3 we present the errors when using LFS values instead of census values.

Table 3

Classification error of the variable "hours worked" in the Labour Force Survey.

Per cent of the census estimate	Hours per week		
	1-19	20-34	35-w
a) Wrongly included	24.6	28.1	6.5
b) Wrongly excluded	49.2	19.8	6.2
c) Gross error	73.8	47.9	12.7
d) Correctly classified	50.8	80.2	93.4
e) Net error	-24.6	8.3	0.2

Remark: Some of the errors come from the fact that the main part of LFS-values refer to another time period.

The results presented in this section do not include the errors shown in Table 3. This does not lead to any serious underestimation of the total error when studying full time employment, but when we estimate other classes of hours work we have to take the classification errors into account. We have also carried out simulation studies on minipopulations where the small areas consist of municipalities. The conclusion is the same; for the most part SYNT-estimator works best of the studied estimators.

Even if we have not used all the available associated variables in the simulation studies, we would still maintain that the SYNT-estimator is the best estimator. It has performed well in the studies and moreover, it has another good feature. Its form is simple.

In the following section we present our work on the refinement of the SYNT-estimator.

4.4 Refinement of the SYNT-estimator

Up to now we have used only the associated variables sex and income but in practice we will even have information about age and industry. In this section we will present the attempt to also include these variables to develop an "optimal" and, over time, robust SYNT-estimator.

The refinement work is possible to carry out without simulation studies because it is

possible to derive the formulas of the (expected) variance and of the SYNT-estimator. For the expected variance, the formulas are the following.

$$E[Var(SYNT)] \doteq \left(\frac{100}{N_{.q}}\right)^2 \sum_{h=1}^{H} N_{hq}^2 \frac{\bar{Y}_{h.}(1-\bar{Y}_{h.})}{nW_h} \cdot K_h , \tag{12}$$

where

$$K_h = 1 + \frac{1-W_h}{nW_h} ; \; W_h = \frac{N_{h.}}{N} ,$$

and for this bias

$$B(SYNT) = \frac{100}{N_{.q}} \sum_{h=1}^{H} N_{hq} (\bar{Y}_{h.} - \bar{Y}_{hq}). \tag{13}$$

Remark: The formula for the expected variance is derived in the same way as in Cochran (1977) Section 5A.8 "Stratification after selection of the sample."

We were interested in using data for all 284 municipalities in Sweden and therefore, the only data source was the Census of Population and Housing. One problem with this approach was that we could not separate persons working the whole year and expected an overestimation of the error. Work with several minipopulations convinced us that even if we were at a high error level we could use data from the census to improve the estimation procedure.

An important feature of an estimator is that it performs well in the long run. As a test of its performance over time, we developed the SYNT-estimator on 1980 census data and then tested it on 1975 census data.

In this study the parameters $\bar{Y}_{h.}$, $\bar{Y}_{hq}$ and N_{hq} are known and the (expected) variance and the bias for a given sample size, n can be computed.

In the following table the work is summarized by a mean value (over municipalities) of rel-RMSE. Rel-RMSE has the following form:

$$\text{Rel-RMSE} = 100 \; . \; \text{RMSE}/T_q, \tag{14}$$

where $\text{RMSE} = \{E[Var(SYNT)] + B^2(SYNT)\}^{\frac{1}{2}}$.

Four associated variables sex, age, income, and industry are categorized in a particular way and denoted SEX, AGE2, INC4 and IND5, respectively. We don't describe in detail the work that led to this categorization.

In practice, the estimator is applied in three consecutive LFS-samples, which aggregates to a sample size of 35,000. To make this study more realistic we used a sample size of 35,000 here too.

Table 4 indicates that we get the best result when we use all of the four associated variables. Excluding AGE2 does not have any significant effect on the mean of rel-RMSE. On the other hand, we can see that SEX is an important associated variable, especially when estimating the percentage economically active persons working more than 34 hours/week.

To estimate the effect of the associated variables we calculated the mean of the rel-RMSE for a synthetic estimator using no associated information (i.e. the nation-wide estimate is used in each municipality). For the class 35-w hours per week and using data from the 1980 census we received the value 2.22, which can be compared with 1.54 for the "optimal" estimator. Where the sample size is 35,000 and we use the DIR-estimator the mean of rel-RMSE is 7.7.

Table 4

Mean value (over municipaliteis) of rel-RMSE
Sample size 35,000

Associated variables	Hours Worked			
	1980 Census		1975 Census	
	20-w	35-w	20-w	35-w
SEX * AGE2 * INC4 * IND5	0.91	1.54	1.07	1.29
SEX * INC4 * IND5	0.92	1.55	1.10	1.34
AGE2 * INC4 * IND5	0.89	1.94	1.05	1.72
INC4 * IND5	0.91	1.90	1.09	1.76

Table 4 shows that the "optimal" SYNT-estimator performs well in both periods studied.

We have also tried several other refinement alternatives, which we describe briefly in the following.

If the bias is stable over time we could use the bias calculated from the 1980 census data and subtract it from the SYNT-estimate in the future. In this study, we make that adjustment of the estimator for the 1975 census and the mean of rel-RMSE becomes for the classes 20-w and 35-w hours/week, 1.07 and 1.12 respectively. Thus, we can not argue that the bias is stable over time.

We also arrived at this conclusion when we tried what is called a SPREE-estimator (Purcell (1979)) using the 1980 census data to describe the association structure.

One way of reducing the bias (model error) in a model-dependent estimator is to group the small areas in such a way that the assumptions underlying the estimators are more likely to be met. In the literature one can see rather successful attempts (Purcell (1979) and Lundström(1987)) to reduce the bias (model error) of a model-dependent estimator by grouping the small areas and calculating the estimates in each group. It is true that the sampling variability will increase, but perhaps not to the same degree as the bias decreases. We have tried to group the municipalities but found that the groupings had no effect after five years.

5. CONCLUDING COMMENTS

The simple synthetic estimator, SYNT, is found to be the best of the estimators studied, but will it provide estimates of acceptable quality? That question has not been answered by our methodological study. Of course, it is difficult to give an answer because it depends on what you expect from these estimators and the trade-offs you are willing to make. When using conventional estimators, at least approximations of the errors are possible to calculate from the sample and thus the statistical users receive adequate information. When using a synthetic estimator the bias is the dominating error and this error can not be estimated from the current data available.

Municipalities are the smallest areas in the methodological study, but there is a need for estimates of much smaller domains. Even for much smaller domains it seems plausible that the SYNT-estimator will be the best estimator.

In spite of all these problems, Statistics Sweden has decided to put the SYNT-estimator into practice.

ACKNOWLEDGEMENTS

I would like to express my thanks to Dr. Claes Cassel for his many helpful comments during the work and to Mr. Göran Råbäck for the statistical computing he undertook for this paper.

REFERENCES

Cassel, C. (1984). Optimal selection of α (in Swedish). Unpublished memo, Statistics Sweden.

Cochran, W.G. (1977). *Sampling Techniques.* Third edition. Wiley, New York.

Hammshek, E.A., and Jackson, J.E. (1977). *Statistical Methods for Social Scientists.* Academic Press, New York.

Hidiroglou, M.A., Morry, M., Dagum, E.B., Rao, J.N.K., and Särndal, C.E.(1984). Evaluation of Alternative Small Area estimators Using Administrative Records. 1984 *Proceedings of the Survey Methodology Section, American Statistical Association.*

Lundström, S. (1986). Estimating Population Characteristics and Households in Swedish Municipalities Using Survey and Register Data. *Proceedings of the Second Annual Research Conference, March 23-26, 1986.* Bureau of the Census.

Lundström, S. (1987). An Evaluation of Small Area Estimation Methods: The case of Estimating the number of Nonmarried Cohabiting Persons in Swedish Municipalities, *Small Area Statistics - An International Symposium.* Wiley, New York.

Purcell, N.J. (1979). Efficient Estimation for Small Domains: A Categorical Data Analysis Approach. Unpublished Ph. D. dissertation. University of Michigan.

Särndal, C.-E. (1984). Design-Consistant Versus Model-Dependent Estimation for Small Domains. *Journal of the American Statistical Association,* Vol. 79, 624-631.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

METHODOLOGY FOR CONSTRUCTION OF ADDRESS REGISTERS USING SEVERAL ADMINISTRATIVE SOURCES

J. DOUGLAS DREW, JOHN ARMSTRONG, ALEX VAN BAAREN and YVES DEGUIRE[1]

ABSTRACT

As part of the research program for the 1991 Census of Population, a study of the feasibility of constructing a dwelling address register for Canadian urban areas is underway at Statistics Canada. The initial pilot test indicated that use of an address register constructed using data from several administrative records systems could improve census coverage. In this paper the methodology used to construct address registers for additional pilot tests scheduled for the fall of 1987 is described. Topics examined include the quality of information available on the various administrative files, procedures used to parse free format address information and record linkage techniques used to unduplicate address lists. The benefits of using information not directly related to address in the linkage process are also considered.

1. INTRODUCTION

The notion of a machine readable household register that could be used in the conduct of population censuses is not new. Indeed, yesterday we heard from Redfern (1987) how not only household registers, but also population registers exist in Sweden, Denmark and some other European countries, and that the existence and use of these registers is in fact reshaping the role of Censuses in these countries. Also, the United States Bureau of the Census uses a list frame of addresses in the conduct of its decennial Census. Private vendor lists form the basis for their list, which is further improved by means of field checks (Whitford 1987).

In Canada, high quality vendor lists do not exist, and so we at Statistics Canada have considered at different times the creation of such a list ourselves. At this point I should note that currently in the Canadian Census, manual address lists are created by some 40,000 Census Representatives, each responsible for an area containing 200-300 dwellings. These lists are created coincident with the drop-off of Census questionnaires, and the address lists are not data captured.

The first study into the feasibility of creating a household or address register was carried out by Fellegi and Krotki (1967). They considered an approach of merging and unduplicating address information from multiple sources — which in their case consisted of the previous Census, municipal assessment roles, and electric utility billing lists. Pilot address registers were constructed and evaluated for two medium sized cities — Waterloo and London. They found the address registers covered 97% of dwellings, which was encouraging. However due to technological limitations of the day, construction of the

[1] Informatics and Methodology, Statistics Canada, 4-C2, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

address registers was largely a manual process, which did not favour implementation at that time.

During the 1970's a series of studies was undertaken which are summarized by Booth (1976). The approach considered was one of data capturing addresses from a previous Census, and using information from Canada Post to update the register and keep it current. The coverage under this approach was found to be comparable to that under traditional census methods. However the high initial data capture costs, despite anticipated savings in the longer term, were viewed as problematic, and the address register was not implemented.

Royce (1986) presented several potential uses and benefits of an address register to Statistics Canada programs, and also enumerated several factors that are now more conducive to construction of an address register than had been the case in earlier decades. These include the increased availability of machine readable administrative record systems with address information, the almost universal use of postal codes on these files, cheaper and more powerful computers, and improved record linkage methods and software. With all of these things in its favour, research into construction and use of an address register in the 1991 Census was started up a little over a year ago, considering an approach to construction like that investigated by Fellegi and Krotki, but with automation of virtually all of the steps. Due to lack of good address information in rural areas, attention is being restricted to urban areas.

Table 1 presents results from a small scale pilot register constructed for an area comprising 5000 dwellings in Ottawa (Drew, Armstrong, and Dibbs 1987). The address register coverage of valid dwellings was found to be about 1% below that of the 1986 Census for the test area. However, it was found that when the Census list and the address register were combined, the resultant list had 2.3% better coverage of dwellings than the Census. It should be noted that the areas chosen for this test were areas of suspected high undercoverage. Census dwelling undercoverage for the test areas was estimated at 3.7% after field verification, so that the 2.3% improvement in dwelling coverage obtained by combining the Census list and the address register represents about 60% of estimated Census dwelling undercoverage. The dwelling overcoverage estimates for Census and address register lists needed to obtain the net dwelling figures involved in Table 1 were obtained using field verification.

Table 1

Ottawa Test: Net Dwellings as % of Census Net Dwellings*

	Dwelling Type		
	Single	Multiple	Total
Address Register	97.9	99.8	99.2
Census + Address Register	102.3	102.2	102.3

*Net Dwellings = Total Dwellings - Dwelling Overcoverage

Based on these encouraging results, a second test was scheduled; this test is currently underway in the field. In this test, two methods for use of an address register in Census data collection are being tested. Both methods are premised on the current drop-off methodology for delivery of questionnaires. The alternative of a mail-out Census based on an address register for 1991 was ruled out early in our research, when a study failed to show it would lead to any cost savings over the traditional methodology, under the assumption that a field check would be required to improve coverage prior to its use (Gamache-O'Leary, Nieman, Dibbs 1986).

Under the first method, which we call the pre-list method, address registers are pre-printed for each Census Enumeration Area. The task of the Census Representative under this method would be to update this list by making deletions and additions as necessary. The updating would be done coincident with the drop-off of Census questionnaires to all valid dwellings.

Under the second method, which we call the post-list method, Census Representatives (CR's) would create address lists from scratch coincident with drop-off of questionnaires, as under the current methodology. After drop-off, the CR would be issued a copy of the address register for his/her Enumeration Area, with instructions to match their manual list against the address register. Any additional dwellings found on the address register would be verified in the field and, if valid, would be added to the CR's list and a Census questionnaire dropped off.

The November 1987 test was restricted to a comparison of dwelling lists under the two methods relative to the traditional Census methodology, and for that reason it did not include any drop-off of questionnaires. For the test, persons with no previous interviewing experience were hired and assigned to a team doing only the pre-list method, or to a team doing only the post-list method. Under the test design, the same areas were listed according to both methods. Persons hired were not aware that two teams existed covering the same areas. As part of the evaluation, we will be data capturing the final address lists obtained under each method and carrying out a computerized match with resolution of any discrepancies through further field work.

This test is being conducted in the Census Metropolitan Areas (CMAs) of Vancouver, Edmonton, Toronto, Montreal and Halifax. For each CMA a stratified sample of 64 Enumeration Areas were chosen, with the stratification being on the basis of predominant dwelling type in the 1986 Census. This sample corresponded to areas of approximately 20,000 dwellings per CMA.

Table 2 presents the administrative files used as sources in constructing address registers for each CMA. Three national files already in Statistics Canada's possession were used for all cities — namely the Revenue Canada personal taxation file (TAX), and Health and Welfare Canada files of Family Allowance (FAM) and Old Age Security (OAS) recipients. In addition, for each site, two lists were purchased from among municipal assessment rolls (MUN), telephone billing lists (TEL) and electric utility billing lists (ELE). Edmonton was an exception in that due to a delay in obtaining one of the extra files, the address register was constructed using only four files.

Table 2

November 1987 Test: Source Files by CMA

CMA	Source File					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	x	x	x	x		x
Edmonton	x	x	x	x		
Toronto	x	x	x	x	x	
Montreal	x	x	x		x	x
Halifax	x	x	x	x	x	

2. STEPS IN ADDRESS REGISTER CONSTRUCTION

As mentioned earlier, the approach to address register construction we are investigating consists of merging and unduplicating address information from multiple administrative data sources. The four principal steps involved are discussed below.

Address Standardization

Address information on administrative files is typically in free format, by which we mean there is no fixed position or even order of appearance for the components of the address, such as street name, street number, apartment number, and so forth. It is necessary to analyse the address information to identify the components, in order that the address can be rewritten in a standard form to facilitate matching. This task turns out to be more complex than one might initially think.

At the outset of the address register research, evaluation studies of existing Statistics Canada software for address standardization revealed sufficient deficiencies that complete redevelopment was felt necessary to support an address register. An expert systems approach has been adopted which incorporates over 100 syntax rules concerning what constitutes a valid address (Deguire 1987). The system breaks the free format address into tokens, which are strings of consecutive letters or numbers, separated by blanks or delimiters such as commas. Some tokens are recognized by the system as keywords. Examples of keywords include 'Street', 'Rue', 'Apt', 'App' and so forth. Based on the pattern of numeric and alphabetic tokens, and known keywords, we have found that it is possible to uniquely decode over 95% of addresses unambiguously into components. While as few as 8 patterns account for 52% of addresses, the number of variations is large and over 1600 patterns are needed to handle 95% of the addresses. Currently the remaining 5% are reviewed and, where possible, deciphered manually. We are concerned about this 5% of cases, and plan to study whether further improvements can be made in the software, and in address register construction what would be the impact of discarding as opposed to attempting manual resolution of such cases.

Merging and Unduplication

After merging the standardized addresses from all the source files, the next step is to eliminate duplicates — that is, records referring to the same address. This is broken into two parts — exact matching to get rid of exact duplicates, and record linkage to identify duplicates where there is disagreement or only partial agreement on one or more of the standardized components. Such discrepancies occur for numerous reasons, such as variations in spelling, use of non-standard abbreviations, and so forth. The record linkage is carried out using Statistics Canada's record linkage software GIRLS (Hill and Pring-Mill 1985), which is based on the Fellegi and Sunter (1969) methodology.

More will be said in the next section about matching and record linkage in relation to construction of pilot registers for the November 1987 test.

Geographic Coding

Since we want ultimately to produce lists of addresses by Census Enumeration Area from the address register, the linkage of the address register to standard census geographic coding at least to the level of Enumeration Area is crucial. This linkage will bear directly on the coverage of the address register at the Enumeration Area level.

A number of possibilities exist for establishing this link and work needs to be done to evaluate them. One means would be through a Postal Code to Enumeration Area link. Such a link was established by data capturing Postal Codes for the one-fifth sample of

dwellings in the 1986 Census, and plans exist for updating and maintaining that link. Plans also exist for evaluating the accuracy of this link, keeping in mind that to use it in linking an address register to Census Enumeration Areas would impose requirements for accuracy and updatedness well beyond what has been needed to support current uses.

Edit and Imputation

The final step in address register construction consists of fine tuning. For instance, logical gaps in apartment numbers can be imputed. Some clearly erroneous addresses which escaped detection at earlier steps in address register construction may be spotted clerically and deleted.

3. PRELIMINARY FINDINGS FROM CONSTRUCTION OF PILOT REGISTERS

In this section, we present some preliminary analysis of the address register construction process, based on the pilot registers for the November 1987 test. More critical and complete analysis will be possible when results of the current field work become available.

Table 3 presents the gross coverage of the pilot registers at various stages of construction as a percentage of 1986 Census dwellings for the test areas. Column (2) indicates the initial number of addresses with Postal Codes corresponding to those in the selected Enumeration Areas in each city according to the most recent version of the Postal Code to Enumeration Area conversion file, whose vintage was February 1987. That is, it represents the number of addresses after merging of standardized addresses from all source files and before elimination of duplicates. The four source files used in Edmonton contained, in total, twice as many addresses as the 1986 Census, while the five files used in other cities contained on average three times as many addresses as the Census.

After elimination of exact duplicates, the gross coverage (compared to the 1986 Census) was brought down from an average of 273% (column 1) to 122% (column 3); this demonstrates the success and importance of the address standardization step.

Table 3

Gross Coverage as % of 1986 Census Dwellings Pilot Address Register at Steps During Construction

CMA	After Merge	After Elimination of Exact Duplicates	After Postal code Verification	After Record Linkage	Final
(1)	(2)	(3)	(4)	(5)	(6)
Vancouver	283	117	109	103	104
Edmonton	194	110	103	99	101
Toronto	283	113	103	102	102
Montreal	312	136	125	111	108
Halifax	297	134	126	109	110
Average	273	122	113	105	105

Column (4) represents a step that was unique to construction of the pilot registers. Postal Codes of addresses were verified using Statistics Canada software designed for this purpose, and cases where Postal Codes were in error and the corrected Postal Codes

fell outside the sample Enumeration Areas were dropped. Note that, in constructing a full scale address register, such cases rather than being dropped would be shifted to the Enumeration Area where they belong. This Postal Code verification step resulted in 9% of the records being dropped with, of course, none being added; it represents a potential source of undercoverage that would be unique to the pilot registers. We plan to assess the extent of such undercoverage, which may range from minimal to being fairly significant depending on the degree of independence of Postal Coding errors from file to file.

The record linkage step reduced the gross coverage by a further 8% (column 5), resulting in average gross coverage of 105%. Column (6) represents the gross coverage after edit and imputation. On average, gross coverage was unaffected, but for individual CMAs it increased or decreased by 1-2%, which is quite a large amount relative to the anticipated net undercoverage of the registers. If results are similar to those for the earlier pilot register for Ottawa, net undercoverage relative to the Census may be close to 1%, which, given the 5% gross overcoverage of the registers, would imply 6% net overcoverage. The overcoverage stems from duplicate records which were undetected in the record linkage process or from appearance on the register of dwellings which are no longer valid.

Results from the field test will tell us the under- and overcoverage not only for the address register, but for its alternative uses in Census data collection. We also plan to do indepth studies of reasons why addresses were missed on the address register, and whether improvements in the methods and software could reduce the undercoverage.

Table 4 presents some results on the record linkage step in address register construction. It presents for pairs matched during record linkage, the percentage of times individual components of the address used in linking either agreed, partially agreed, or disagreed. It should be noted that street number was a blocking factor in record linkage, that is searches for links took place only amongst records which agreed on street number. Another point worth noting was that during the merge and exact matching, a record was kept of the source files on which each address appeared, and during record linkage it was the version appearing on the most source files that was retained. Two levels of partial agreement were allowed as comparison outcomes for street and municipality names. (These are combined in Table 4). The first level consisted of cases of minor misspellings due to omission of a letter or transposition of two letters. Two names were declared to agree at the second level of partial agreement if their phonetic versions coded using the NYSIIS (New York State Identification and Intelligence System) scheme were identical. NYSIIS coding is intended to eliminate the effects of common spelling errors.

Table 4

Comparison Rule Outcomes for Address Pairs Matched by Record Linkage (Percentages)

Matching Category	Outcomes			
	Agreement	Partial Agreement	Dis-Agreement	Missing
Street Name	49	31	20	
Apt. Number	93		7	
Civic Number Suffix	95		5	
Postal Code: Dig. 1-3	100			
Postal Code: Dig. 4-6	95	4	1	
Municipality	87	2	11	
Family Name	35		18	47

Another field where partial agreement was allowed was in the last three characters of the Postal Code, where two out of the last three characters being the same constituted partial agreement.

It is interesting to note the low frequency with which the street name agreed for matched records, with full agreement only half of the time. This appears to be due to frequent misspellings and abbreviations. Another point worth noting regards the use of family name as a match variable. This variable was used only for record linkage purposes, and was deleted from the final register. Due to the different ages of the source files, failure to link on family name was not counted against linking a pair of addresses; however, agreement on name was considered quite important, that is it received a high positive weight. In order to assess the impact of using family name, for one city we repeated the record linkage without name, and found that 1% less records were linked.

The next two tables examine the contributions of the various files to the final address register. Table 5 presents coverage of the source files as a percentage of address register gross coverage — that is, what percentage of the address register records were traceable back to each of the source files. This table confirms as we had suspected that coverage of the tax, telephone and electric utility files is high. The electric utility files came out best, and it appears, at least in the two provinces we have looked at, that bulk metering of multi-unit structures, which previously had been a weakness of this source is no longer a significant factor. The low tax file coverage in Montreal and Toronto was due to frequent errors in the tax file address leading to its not being the retained version. The coverage of the municipal assessment file, except for Toronto, was quite low since they generally have only one record per owner for multiple unit structures.

Table 5

Gross Coverage of Sources Files
(% of Address Register Gross Coverage)

City	Source File					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	73	26	26	48		87
Edmonton	82	32	18	49		
Toronto	60	22	18	78	76	
Montreal	57	24	16		72	86
Halifax	78	30	19	47	72	

Table 6 gives the percentage of addresses uniquely contributed by each source. Once again, electricity files performed very strongly, and the telephone files were not far

Table 6

Unique Contribution by Source File
(% of Address Register Gross Coverage)

City	Source File					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	5	1	1	1		13
Edmonton	28	5	4	4		
Toronto	2	0.5	0.5	7	12	
Montreal	3	0.5	1		9	17
Halifax	10	1	1	2	9	

behind. The tax files performed well in the case of Halifax and in Edmonton. The Edmonton result is anomalous in the sense that of the four files used, the tax file was the only one with high coverage of addresses.

It should also be noted that these results are for the contribution of individual files to gross coverage. It will be of interest, once the field results from the November 1987 test are available, to see the contribution of each file to net coverage. The usefulness of files such as Family Allowance and Old Age Security would be very questionable if a substantial proportion of the 0.5-1% unique addresses they contribute are in fact found to be in error.

Future Directions

Analysis of the results from the November 1987 test will be completed by the spring of 1988. Also estimation of the developmental requirements, and cost and timing implications of different scenarios for use of an address register in the 1991 Census will be completed by that time. A decision on the extent of use of an address register in the 1991 Census, based on these two inputs, is scheduled for the spring of 1988. If a decision to use an address register on a wide scale is taken, this will imply a high priority to developmental work leading up to 1991.

The work to date has identified areas for further research, some of which would have to proceed in parallel with development if the decision taken is in favour of implementation. The research should continue also if it were decided to use an address register on a test as opposed to a production basis in 1991. Areas where further research is needed are discussed below.

Updating Methodology

To date we have only considered the initial creation of an address register. The sources and approaches that are best for creation are not necessarily the best for updating. Consideration has to be given both to the frequency with which updating is needed, and the implications on systems design of the frequency and proportion of updates. One possible approach to updating would be to do an exact match on successive versions of source files to identify changes, which would then be linked to the existing address register. The handling of deletions of addresses that are no longer valid under such an approach needs special investigation. Another possibility would be the use of data sources such as construction and demolition permits or updates from Canada Post.

Use of Address Register in Enumeration Area Delineation

For collection and dissemination purposes, Census Enumeration Areas should contain approximately the same number of dwellings, and they must respect higher level geo-statistical and geo-political boundaries. Since dwelling counts used in Enumeration Area delineation are currently primarily based on the previous Census, they are sometimes quite out of date. Dwelling counts from an intercensally updated address register should improve the delineation process, and reduce the expense and disruption of having to split Enumeration Areas due to discovery of substantial growth during field operations for the Census.

Use of the Address Register as a Frame for Household Surveys

Currently most household surveys at Statistics Canada are based on area samples, which require costly face to face interviewing, at least in the first month households are sampled. Telephone frames by themselves are not a viable alternative for large national surveys, due to the bias associated with undercoverage of the non-telephone universe

(Drew and Jaworski 1986). The alternative of dual frame methodologies combining area frames and telephone frames is fairly inefficient in the sense that a relatively large area sample is needed to cover the small non-telephone population. An address register with telephone numbers for roughly 75% of urban households (see Table 5) has appeal as a frame which can afford the benefits of telephone interviews for a large portion of the urban population, while identifying and permitting the adoption of an efficient sample design for remaining urban and rural households.

Plans are to convert a portion of the Labour Force Survey sample to an address register based design in the areas where pilot registers are being maintained for use in Enumeration Area delineation. As part of the test, methods for dealing with address register undercoverage will be investigated.

Refinement of Address Register Methodology

Finally research is needed into the address register construction process itself. Issues such as the impact of more or different source files need study. Can additional sources with high coverage be found, and if so what would be the implications of their use in address register construction?

Also, we saw that the software for address standardization, and for validating Postal Codes does not successfully handle all cases. More needs to be known about the problem cases. Are they cases of address errors appearing on one file while valid versions of the same address appear on another file? If this were the case, ignoring problem cases on individual files might be the recommended course of action. If cases not handled by the software are due to systematic failure of the software to handle valid addresses, or if particular addresses tend to be in error on all files, then ignoring these cases would lead to coverage problems. Detailed study of problem cases, including addresses missed on the pilot registers is needed to answer these questions.

In summary, the findings to date are encouraging, both in terms of the technical feasibility of producing at a reasonable cost an address register with high coverage of urban addresses, and in terms of the potential for such a register to reduce undercoverage in the Census. A number of avenues of further research into uses and improvements of the methodology for construction and updating are planned for the coming year, to be integrated with developmental work should the decision be taken to proceed with implementation of an address register for the 1991 Census.

REFERENCES

Booth, J.K. (1976). A Summary Report of All Address Register Studies to date. Internal Report, Statistics Canada.

Deguire, Y. (1987). Research into the Parsing and Standardization of Free Format Addresses at Statistics Canada. Internal Report, Statistics Canada.

Drew, J.D., Armstrong, J., and Dibbs, R. (1987). Research into a Register of Residential Addresses for Urban Areas of Canada. *Proceedings of American Statistical Association, Section on Survey Research Methods.*

Drew, J.D., and Jaworski, R. (1986). Telephone Survey Development on the Canadian Labour Force Survey, *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Fellegi, I.P. and Krotki, K.P. (1967). The testing programme for the 1971 Census in Canada. *Proceedings of the American Statistical Association, Social Statistics Section, 29-38.*

Fellegi, I.P., and Sunter, A.B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Gamache-O'Leary, V., Nieman, L., and Dibbs, R. (1987). Cost Implications of Mail-out of Census Questionnaires using an Address Register. Internal Report, Statistics Canada.

Hill, T., and Pring-Mill, F. (1985). Generalized Iterative Record Linkage System. *Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, 327-333.

Redfern, P. (1987). European Experience of Using Administrative Data for Censuses of Population: The Policy Issues that Must be Addressed. International Symposium on Statistical Uses of Administrative Data, Ottawa.

Royce, D. (1986). Address Register Research for the 1991 Census of Canada. Journal of Official Statistics, 2, 447-456.

Whitford, D. (1987). Research Program for the 1990 Decennial Census, *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

MULTIPLE USES IN STATISTICS OF ADMINISTRATIVE RECORDS IN THE ANALYSIS OF EDUCATION DATA

CHARLES D. COWAN and MARY K. BATCHER[1]

ABSTRACT

Education data presents a particular analytical challenge. The data available are often hierarchical, meaning that there are several levels at which one can collect, summarize, and analyze the data. While this situation is not unique to education, what is unique is the variety of data available and the opportunities for horizontal and vertical integration of survey and administrative data. The first part of this paper will review current uses of administrative data by the Center for Education Statistics. At the elementary and secondary level, administrative data on school districts, public schools, and public school teachers are collected and summarized for each state and the U.S. as a whole. One use of this data is compilation and presentation as summary statistics. A second use of the data is analysis of the data over time to determine trends in resource availability and usage.

The second half of this paper will preview two other uses of the administrative data. A third use of this data will be to supply population values to be used in weighting of the School and Staffing Survey, a large scale survey of school districts, schools, and teachers to be conducted for the Center for Education Statistics starting in 1988. A fourth use of the data will be to create and validate forecasting models for enrollment, demand for teachers at the elementary and secondary levels, teacher to pupil ratios, and graduation rates from high school on a state by state basis.

Parallel uses of administrative record data for postsecondary education, like transcript data from colleges and universities, will also be discussed. The integration of these data with survey data collected by the Center help reduce both bias and variance in estimates of trends and levels in the U.S. education system.

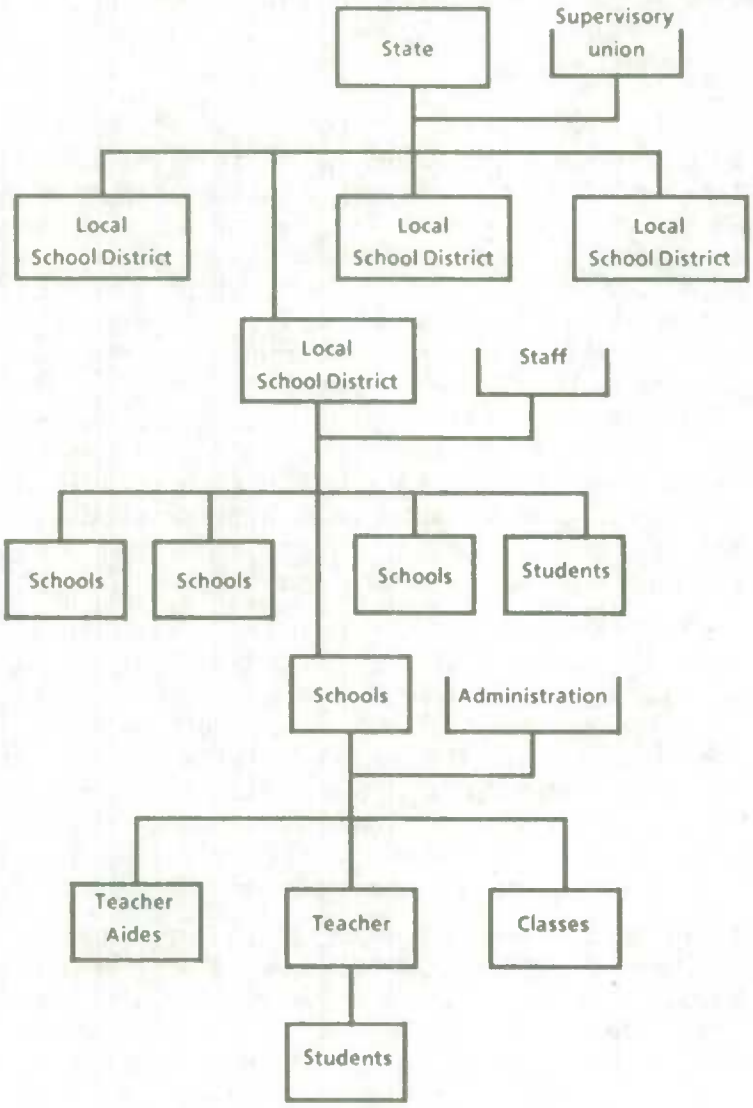
1. INTRODUCTION

Education data are different from data collected from the general population in that the data come from multiple levels, each level consisting of administrative units which are to be studied or which provide data regarding lower levels. Chart 1 displays the different levels from which records come in the U.S. Educational System. These

[1] Charles D. Cowan and Mary K. Batcher, Center for Education Statistics, 555 New Jersey Avenue North West Room 400, Washington, D.C. 20208. U.S.A.

relationships lead to some unique problems in the collection, validation, and analysis of the administrative data. This paper will describe how education data are collected and used in administrative record systems, and the broader issues faced in linking the administrative data with survey and census data collected by the Center for Education Statistics (CES).

CHART 1: LEVELS OF EDUCATION DATA



2. ADMINISTRATIVE RECORD SYSTEMS

There are two major systems of administrative records currently collected within CES: the Common Core of Data (CCD), which addresses elementary and secondary education concerns, and the Integrated Postsecondary Education Data System (IPEDS), which collects data on postsecondary education in the United States. Both systems will be described in this section, and then the remainder of the paper will be devoted to reviewing uses of the administrative record systems for estimation and analysis.

2.1 The Common Core of Data

The Common Core of Data is an interrelated set of surveys of State Education Agencies in the 50 States, the District of Columbia, and outlying areas under U.S. jurisdiction. This collection provides CES with basic information about the approximately 85,000 public schools in the U.S., their 40 million students, over 2 million teachers, and 4 million total staff. It also includes detailed reporting of $137 billion of annual expenditures for U.S. public schools. The CCD is an annual collection conducted continuously since fall 1977.

While the U.S. Department of Education has collected basic information about U.S. public schools since its formation, the immediate predecessor of CCD was the Elementary and Secondary General Information Survey (ELSEGIS), initiated in 1967. ELSEGIS was designed to complement program-specific information from Federal program data systems with general information about school systems and students. These instruments collected data on the number of schools, their grade levels, enrollments and pupil-teacher ratios in local school systems. ELSEGIS also collected information on the universe of schools, including school name, enrollment, and number of teachers. The items in ELSEGIS are similar to those in CCD.

The current CCD collection was pilot-tested in 1976 and implemented in the 1977-78 school year. CCD is now undergoing a three year process of evaluation and revision as part of a joint Federal/State effort to improve the database.

The CCD is presently made up of 4 surveys: the State Nonfiscal Survey, the Survey of Revenues and Current Expenditures for Public Elementary and Secondary Education, the Public Elementary/Secondary Education Agency Universe Survey, and the Public Elementary/Secondary School Universe Survey. All of the CCD surveys collect administrative record data from State Education Agencies. The State Nonfiscal Survey collects numbers of students by grade, teachers, administrators, guidance counselors, librarians, and other staff, and high school graduates. The Survey of Revenues and Current Expenditures for Public Elementary and Secondary Education collects, in addition to revenues and current expenditures, the components used in calculating average daily attendance, for public elementary and secondary schools in the U.S.. The Survey of Revenues and Current Expenditures collects data for the prior school year and is thus lagged one year behind the other survey components of CCD. All data collected on both the State Nonfiscal Survey and the Survey of Revenues and Current Expenditures are reported to CES for the State as a whole.

The other two parts of the CCD collect data at a less aggregated level. The Public Elementary/Secondary Education Agency Universe collects name and address, organizational type of agency, metropolitan status of the area served, number of schools and their grade span, and numbers of students instructed, for each school district by State. The Public Elementary/Secondary School Universe Survey collects name and address, school type, and student and teacher counts for all public schools in every State.

Beginning in August 1985, CES funded a three-year project with the Council of Chief State School Officers to describe the current status of data contained in the CCD, to explore the addition of data elements to make the database adequate and appropriate for reporting on the condition of the nation's schools, and to prepare recommendations to States and the Center for making the database more comprehensive, comparable, and timely. Recommendations have now been received by CES and final decisions made on all nonfiscal data elements except staffing; changes to the nonfiscal collections, excluding staffing, will be implemented in the 1987-88 collection. Recommendations on staffing and fiscal components will be received and evaluated by CES and any resultant changes will be put in place over the next two years (1988-89 and 1989-90 school years).

The new nonfiscal data elements to be added in 1987-88 include telephone numbers for schools and school districts, urban status of schools, student counts, at the school level, by racial category and eligibility for free lunch programs. The school district universe counts will add more detailed student counts and incorporate, for the first time at the district level, counts of high school completers.

Other measures taken to improve the database include the standardization of definitions across States and increased training for the data providers, the CCD coordinators in each State. This standardization of definitions has led to some breaks in the time series. These breaks result from two conflicting needs: (1) the need to standardize definitions and procedures to make the data comparable across States, and (2) the need to provide comparable data over time. When the revisions are complete and have been in place for a few years, it is hoped that these two needs will no longer be in conflict. However, until the time series stabilizes, it will be necessary for CES to map from the old to the new where possible and to provide explanations when a precise mapping is not possible.

In prior years, CCD was operated in a somewhat independent fashion, with the different parts mailed out separately, on different dates, and with different due dates. Beginning in 1987-88, the parts of CCD will be integrated into a single package with a mailout at the end of December and the same due date of March 15.

For the nonfiscal data, preliminary reports will be published in June with data reported as of May 15, for the school year that is then just ending. The Center will accept revisions from States through September 15 and publish final data in October for what will then be the previous school year. The nonfiscal files will be opened to accept a single round of late revisions in September of the following year and will then close and not be reopened. Final fiscal data will also be published in October but will be lagged one year behind the nonfiscal. When the nonfiscal data files close in September, tapes of the final and previous year's revised final data will be prepared and made available for purchase. Tapes of the fiscal data are also available, but they are on a slightly different schedule. For the State aggregate tapes, several years of data will be included on the tapes.

2.2 The Integrated Postsecondary Education Data System

IPEDS is a system of surveys designed to collect data from all primary providers of postsecondary education. The areas covered in IPEDS include institutional characteristics, enrollment, completions, finance, staff, and salaries. IPEDS is intended to be used to report on the condition of postsecondary education in the United States; to do this, CES must have a system to describe postsecondary education and to follow changes in its size, character, providers, and participants.

In the past, CES accomplished this through three major surveys. The Higher Education General Information Survey (HEGIS), The Vocational Education Data System (VEDS), and The Survey of Noncollegiate Postsecondary Schools with Occupational Programs. The data from these surveys have been supplemented by "special studies", including periodic HEGIS surveys (e.g., Residence and Migration of College Students, Surveys of College and University Libraries); a sample survey of Recent College Graduates, Fast Response Surveys, and collaborative surveys with other Federal agencies. Through the development, implementation and operation of these surveys, CES became aware of several inherent methodological problems. For example, because of considerable overlap in the data collection universes associated with HEGIS and VEDS, institutions involved in both of these collection efforts were confronted with extra data burden. Even so, CES could not synthesize the extra information due to differences in data defintions, survey procedures, and the like. In addition, several segments of a larger universe of providers of postsecondary education were not included or even identified in these data collection efforts. Thus, a complete description of the postsecondary education enterprise could not be provided by this survey program.

In recognition of these problems and based upon recommendations from the postsecondary education community, CES developed the Integrated Postsecondary Education Data System over a three-year period. IPEDS encompasses all providers of postsecondary education and permits a complete and adequate description of the postsecondary education enterprise. Throughout its development, every step was taken to ensure that IPEDS meets the following objectives:

- Eliminate duplication and redundancy in postsecondary education data collection;
- Minimize data burden;
- Permit similar data to be comparable across postsecondary sectors;
- Allow for the unique factors of the different postsecondary sectors; and
- Provide valid and reliable statistics from postsecondary education providers.

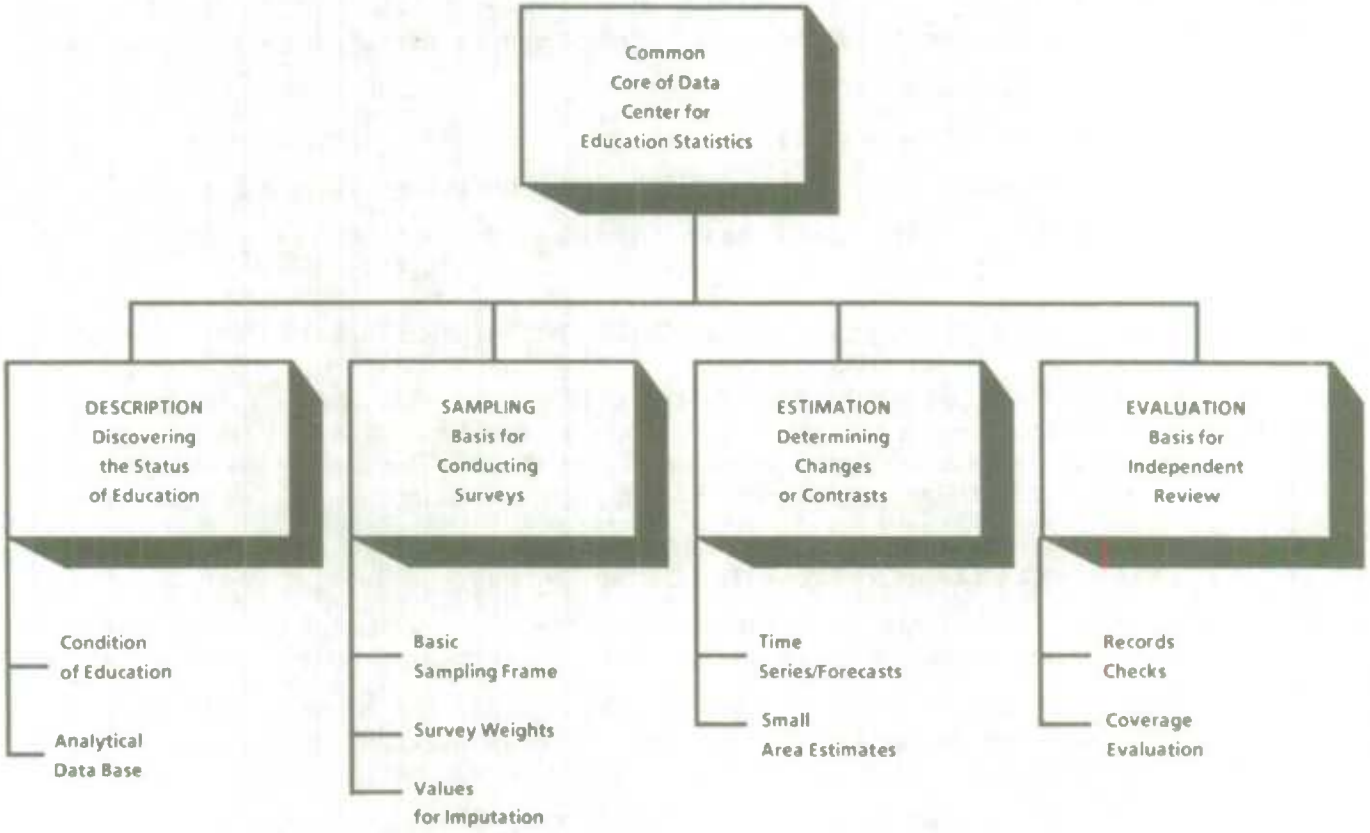
3. USES OF ADMINISTRATIVE RECORDS, WITH EMPHASIS ON CCD

The remainder of the paper will describe the wide variety of uses which can be made of administrative records with respect to the reporting functions of a statistical agency. To illustrate these, the discussion will focus on uses of CCD. The uses of administrative records for a Federal statistical agency can be summarized in eight general categories:

1. The data are used to summarize and describe the current distribution of resources and the condition of education in the United States;
2. The data are used as part of a time series to describe trends and forecast certain key indicators describing where education will go in the next year or five years;
3. The data can be used as a sampling frame for stratification or selection of schools or districts with probability proportionate to some measure of size available in the administrative records;
4. The data can be used to provide totals and margins that will form the basis of sample weighting to reduce the variance and selection biases in Center surveys;
5. The data can be used to make small area estimates of factors collected on a sample basis;
6. The data can be used in the imputation of missing data in sample surveys; and,
7. The data base is made available for other researchers to query for analyses they may wish to conduct.
8. The data can be used to evaluate surveys and censuses for both coverage and validity of response.

Chart 2 displays the various uses of administrative records.

CHART 2: USES OF ADMINISTRATIVE RECORDS FOR EDUCATION DATA



3.1 Reporting the Current Condition of Education

The primary use of CCD data is to report on the condition of U.S. public elementary and secondary education. It is thus a major source of data for CES publications. CCD provides much of the data for the two key annual CES publications, the Digest of Education Statistics and the Condition of Education as well as a number of short reports and other special publications.

Part of the reporting function of CCD is its role as the major determinant in allocations formulas for some Federal programs. The per pupil expenditures for each State are calculated using information reported to CES on the Survey of Revenues and Current Expenditures for Public Elementary and Secondary Education. Certain funds, distributed under various programs, including Chapter 1 of the Education and Consolidation Act of 1981, Impact Aid, and Indian Education, are based, in part, on the State per pupil expenditure.

3.2 Forecasting Models

CCD is used as the basis for the Center's annual projections of enrollment, demand for teachers at the elementary and secondary levels, teacher to pupil ratios, and graduation rates from high school. Most recently, the forecasting uses have expanded in the area of elementary and secondary education by developing State forecasts of enrollment and numbers of teachers. These models are more complicated than the traditional national estimates because of movement of the population and differences between States in certain basic rates related to retention in school. The first such estimates produced this year required CCD data over the past decade by State, State level summary data from the Bureau of the Census, State level forecasts of the population between the ages of 4 and 17, also from the Bureau of the Census, and finally calculation of grade retention rates including the flow of students in and out of private schools.

An interesting unexpected benefit of the development of the forecasting models was the ability to use the models for each State to validate the most recent reports of enrollment from each State. In the development of the time series, preliminary reports from States were used as the final data point in the series. But in some States it was obvious that some of the reported enrollments fell outside the expected range, given the forecasting model. Nine States were contacted to verify the information received, and in two cases additional data were provided, which strengthened the model. In the other States, the seemingly anamolous data points were explained by changes in State reporting or the introduction of new legislation in the State which affected education in that State.

3.3 Administrative Records as a Sampling Frame

Another important use of CCD data is as a sampling frame for surveys of U.S. public schools and school districts. Both the Public Elementary/Secondary Education Agency Universe and The Public Elementary/Secondary School Universe include name, address, some additional descriptive information, and enrollment for all schools and school districts in the nation. In addition, the Public School Universe includes counts of classroom teachers for all schools. These lists have been frequently used as a sampling frame for surveys of public schools and school districts. One recent example was the Center's 1985 Public School Survey, in which a nationally representative sample of 2801 schools was drawn from the CCD Universe lists.

The use of CCD as a sampling frame was an important consideration in the addition of new data elements to these lists, such as counts of students eligible for free lunch programs and counts by racial/ethnic categories. The addition of new data elements, faster processing of the school and education agency universe lists, and close work with States to make these lists complete should make them more useful sampling frames. There is no similar list for private schools, so CES is in the process of creating one with assistance from the Bureau of the Census. The process of creating lists has three parts. The first is the use of a purchased list from a corporation in the U.S. which specializes in education data. This list was initially constructed for use by book publishing companies as a potential market list. The second part of the process is to get administrative record lists from a large number of associations of private schools. These associations are religious or secular and cover a large proportion of the larger private schools. The third part of the frame construction process is to build an area frame in approximately 75 PSU's, which are thoroughly canvassed to turn up any additional private schools not found on the first two lists. When this process is complete, CES expects the private school frame to be 95 percent complete, with most of the undercoverage being smaller schools.

3.4 Weighting Sample Data

Another important potential use of administrative record systems which are complete is their use for weighting of sample data to known, fixed population parameters. Such weighting is valuable in three respects. The first is that it reduces the variance of the sample estimates, since the adjustment incorporates data with no sampling variance, the record system being complete. The second value is that it may remove selection bias that comes about in the sampling process. The distribution of cases in the sample is shifted about to emulate the known distribution in the population. Thirdly, published estimates from the surveys are consistent with other published figures from the administrative record system. This weighting adjustment is commonly used in general population surveys, where the final stage of the weighting process is to rake the survey distribution to a known (or estimated) population total. This type of adjustment can also ameliorate undercoverage in the sampling frame if the researcher using the administrative record system believes the record system, or a modified record system to allow for undercoverage, can provide better estimates of totals.

3.5 Using Administrative Records for Small Area Estimates

There are times that the researcher wants to have information at a very low geographic or demographic level, but does not have the resources necessary to conduct a sufficiently rigorous study at the desired level of analysis. In this case the administrative record system may be of use when combined with the survey data to create estimates at lower levels of analysis. There are two ways in which the administrative record system can be used. The first is direct synthetic estimates based on survey relationships applied to the administrative record system. The relationships may be estimated at the State level (like average level of resource availability as a function of number of teachers and enrollment), and then applied to administrative records for an estimate at the school district level. The term "synthetic" is appropriate here, because the relationships are assumed to hold at all levels of analysis, and no use is made of the data available from the survey at lower levels of geography.

The second way to use the administrative record data is to develop estimates at the desired level of analysis using borrowed strength estimators. With this technique, relationships are modelled, and these relationships are used to make estimates for areas where no sample data was collected. This is an especially useful technique for education studies since borrowed strength can use data from multiple levels (e.g. region, State, school district, and ultimately school) to make estimates at the lowest level, with fairly strong estimates in the model at higher levels, and correspondingly less strength or information at the lowest levels.

3.6 Imputation of Missing Data

Imputation for nonresponse in sample surveys has commonly used a "hot deck" procedure. An alternative to this approach is the use of information and relationships known from administrative record data. This alternative can be used to directly fill in a missing value when the population and specifically the case receiving the imputation is known to be fairly stable. When this assumption fails, but a relationship is known that would inform the imputation, past or current values from CCD can be used to form the base of an estimate for imputation, with random or nonrandom perturbations used to construct a final imputed value.

3.7 Availability of the Data Base to Other Users

The data bases developed for the Center are also set up as public use files and are sold for a nominal fee to any researcher desiring the information. In addition, extensive codebooks and examples of the proper use of the data for analysis are prepared and available as documentation to accompany the tapes. Researchers at other Federal agencies, at universities, and in associations in the private sector have made use of data from CCD for analysis and description of the education system in the U.S.. The CCD data are used by other components of the Department of Education in the annual report of the Secretary of Education, by the Bureau of the Census in their calculation of migration rates, and by organizations like the National Governors Association and the Council of Chief State School Officers in development of indicators on the quality and progress of education.

4. QUALITY ISSUES IN THE COMPILATION OF THE ADMINISTRATIVE RECORD SYSTEM

Quality issues focus on three broad areas; they are:

1. comparability of the data across survey forms,
2. comparability of the data across States, and
3. comparability of the data over time.

With respect to comparability across survey forms, there are a few key data items that are repeated on several forms. Enrollment is collected at State, local school district, and school levels. Recently, CES aggregated the enrollment reported at the school level to obtain State totals. For most States, that total was not equal to the enrollment reported on the State Nonfiscal Report. Differences tended to be quite small, but the existance of differences on universe collections is troublesome in itself. Some of the differences stem from definitional problems and a lack of specificity.

Historically, most States have included on the State Nonfiscal Report only those schools that they support and over which they have some jurisdiction. They are, however, willing to include, on the School Universe Report, other schools that are operated using public funds but are not under the control of the State Education Agency. Thus, schools operated by other State agencies, like youth correctional institutions, may be included on the School Universe Report but not on the State Nonfiscal Report. Conversely, in some States, there are students who are assigned to programs but not to schools and are thus included on the State Nonfiscal Report but not on the School Universe. These differences, although small and generally explainable with some detective work, have remained an unresolved problem.

As CES works with States to improve its database by making it more useful to policymakers and researchers, the focus has been on refining the data elements collected and on making the data collection practices and definitions uniform. Although CES has a long history of publishing handbooks and definitions, recently an intensified effort has been made with States to clarify definitions and to make collection and reporting as similar as possible across States. Negotiations were conducted one-on-one with States to determine exactly what States can and cannot provide and to identify any differences between what the CCD collection requests and what States can actually provide to CES. Where differences were identified, they were carefully spelled out and efforts were made to identify the measurement implications of those differences. As a part of the negotiations, States were asked when they could provide the requested data according to CES definitions. There has thus been a concerted effort on the part of the Center to

move the States toward more uniform definitions and reporting and to make the data collected more comparable across States.

As States work jointly with the Center and move toward more uniform definitions and procedures, some of those changes will cause discontinuities in data reported to CES. As stated earlier, this poses a problem for any trend analysis or projections that rely on time series data. It is anticipated that, while the move toward uniformity across States has perturbed the time series in some States and will continue to cause some problems for a few years, these are not insurmountable difficulties. Adjustments can be made to the projections to accomodate these differences and discontinuities. With continued effort on the part of CES the definitions and procedures will stabilize again but with a greater degree of uniformity and comparability across States.

5. SUMMARY

The statistical uses of administrative record data are many and varied. Administrative record data are an important component of the Center's series of products. These data are used descriptively in reporting on the condition of U.S. education, in forecasting, as a sampling frame, in sample weighting for variance and bias reduction, in small area estimation, in the imputation of missing data, and as a database for use by researchers outside the Center. The validity of the results of these statistical uses is dependent upon the comparability of the data over time, across forms and across States. The Center is engaged in a process to assess and improve the quality of this valuable resource.

6. REFERENCES

Education in the States. (1987). Volume I: State Education Indicators. Council of Chief State School Officers, Washington, D.C.

Results in Education: (1987). National Governors' Association, Washington, D.C.

Magnani, Robert J., Cowan, Charles D., Biemer, Paul P., and Turner, Anthony G.. (1985). Evaluating Censuses of Population and Housing. Statistical Trianing Document ISP-TR-5, Bureau of the Census, Washington, D.C., September 1985.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

THE SOCIAL POLICY SIMULATION DATABASE
AN EXAMPLE OF SURVEY AND ADMINISTRATIVE DATA INTEGRATION

MICHAEL WOLFSON, STEVEN GRIBBLE, MICHAEL BORDT,
BRIAN MURPHY and GEOFF ROWE[1]

ABSTRACT

This paper describes the construction of a prototype database explicitly designed to support analysis of personal income and sales tax and income transfer policies. Tax and transfer policies increasingly require integrated analysis that cuts across traditional jurisdictional and program lines. The Social Policy Simulation Database/Model (SPSD/M) was constructed to support micro-analytic modeling by combining individual administrative data from personal income tax returns and unemployment claimant histories with survey data on family incomes and expenditure patterns. Considerable use of additional aggregate administrative data was made in both the database creation and modeling phases of the project. Input-output data were also applied in modeling sales taxes and duties as they relate to personal consumption. The techniques used to create the database and avoid confidential data disclosure include various forms of stochastic matching and imputation.

1. INTRODUCTION

Many current issues and problems in public policy focus on the economic positions of individuals and the families within which they reside. These include questions relating to personal taxes, unemployment insurance benefits, and social welfare payments. The focus is not only on the aggregate figures such as total taxes paid and average pension benefits, but also on such detailed distributional matters as the pattern of transfer program benefits by income group, and the composition of types of families, for example, in the lowest income ranges. The analysis of such questions requires more than tables of statistics. It requires microsimulation: the effective use of modeling techniques on a representative sample of individuals and families containing a broad range of data.

Microdata collected by different statistical surveys and administrative procedures are designed for specific purposes. No one dataset provides a sufficiently integrated and detailed picture of Canadian households to support the analysis of personal tax and transfer policy issues. For example, unemployment insurance administrative statistics, are drawn from an individual-based weekly-wage-oriented program, and thus contain no information on the families to which the claimants belong nor on any other sources of

[1] Michael Wolfson, Steven Gribble, Michael Bordt, Brian Murphy and Geoff Rowe, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario. K1A 0T6.

income they may have. Although unemployment insurance is an individual-based program, it is still desirable to analyze the impacts of change in UI regulations on family incomes and on other social programs such as social assistance and personal income tax revenues.

The Social Policy Simulation Database with its related Social Policy Simulation Model software (SPSD/M) has as its general goal to provide a comprehensive, publicly available, microsimulation-based, integrated individual tax/transfer policy analysis capability. This paper gives an overview of the construction of the microdatabase, the SPSD.

The paper first describes the general objectives of the SPSD and character of the source data. Then, in the main part of the paper, the many steps in the assembly of the SPSD are described.

2. OBJECTIVES, DATA SOURCES, AND TECHNIQUES

In developing the SPSD every attempt has been made to maintain the variety and utility of the original source data while ensuring the confidentiality of these data so that the resultant database and model can be publicly released. Four central objectives thus guided the selection of techniques, data sources and variables, and process:

- **Public Accessibility/Non-Confidentiality**

  The first objective has been to ensure that no actual individual represented in any of the databases could be identified through either explicit or residual disclosure. Only with this requirement met can the SPSD/M be released to the public in an effort to improve the richness of public analysis and debate in the social and tax policy areas in Canada. Also related to public accessibility is the requirement that the database and model be capable of executing on a moderately priced microcomputer with limited storage facilities.

- **Aggregate and Distributional Accuracy**

  The SPSD/M has been designed to reproduce as closely as possible "known" aggregate totals such as total UI beneficiaries. At a secondary level attempts have been made to represent the distribution of aggregates across several classifications key to public policy analysis in Canada such as province, age, income, family type, and sex. Subsequent distributions have been preserved where possible.

- **Completeness and Detail of Data**

  The selection and aggregation of variables from the main data sources has attempted to foresee likely policy options as well as serve the needs of the current tax/transfer models. For example child care costs are included in the database yet are not currently used in any of the models.

- **Micro-Record Consistency**

  Wherever possible consideration has been given to avoiding the creation of unrealistic individual persons and households. For example an elderly childless couple with a full child care expense deduction.

These central objectives are highly interdependent and compromises among them have been made. The process of making trade-offs involved participants from several federal departments with an interest in the results as well as previous experience with their own simulation models. The ultimate result then represents a compromise between methodological, informational, technological, departmental and public policy concerns.

The SPSD has been constructed from four major sources of data.

- ***The Survey of Consumer Finances (SCF):*** Statistics Canada's main source of data on the distribution of income amongst individuals and families served as the host dataset. It is rich in family structure and income details; but it lacks detailed information on unemployment history, tax deductions and consumer expenditures.
- personal income tax return data: the three percent sample of T1 returns used as the basis of Revenue Canada's annual ***Taxation Statistics*** (Green Book) publication;
- unemployment insurance (UI) claim histories: a specially drawn one percent sample of histories from the EIC administrative system; and
- the ***Family Expenditure Survey (FAMEX):*** Statistics Canada's periodic survey of very detailed data on Canadian income and expenditure patterns at the household level including information on net changes in assets and liabilities (annual savings).

These original data sources are confidential and have not been released in their complete form. Instead, they are disseminated either as public-use samples in which many records and variables are suppressed, or in the form of summary tables, individual cells of which may be suppressed for confidentiality purposes.

For purposes of the Social Policy Simulation Database (SPSD), these four data sources have been transformed into a single non-confidential public use microdataset. In addition, these microdata have been augmented by reference to various aggregate data which served mainly to provide benchmarks or control totals. These aggregate data were drawn from the 1981 Census, Canada Assistance Plan administrative reports, Statistics Canada's 1981 census, Vital Statistics, Health and welfare summary reports as well as the microdata sources themselves.

The joining together of the four initial microdatasets, addition of new information and the replacement or adjustment of biased measures were largely dependent on four techniques employed extensively in the creation of the SPSD: iterative proportional adjustment, stochastic imputation, micro-record aggregation, and stochastic merging.

Iterative proportional adjustment (IPA) refers to a technique for reduction of bias by forcing agreement between data and known control totals. For example, survey weights may be adjusted to ensure that the population by age and sex represented by the survey corresponds to the "known" population by age and sex (i.e. based on census data).

Stochastic imputation is the generation of synthetic data values for individuals on a host data set by randomly drawing from distributions or density functions derived from a source data set.

Micro Record aggregation is the process of creating synthetic micro-records by statistically combining groups of similar records. Micro records on the host dataset are grouped according to policy-relevant criteria. Within each group values of relevant variables (e.g. capital gains) are averaged to create non-identifiable records which resemble microdata but are actually synthetic.

Stochastic merging involves first classifying records on both a host and donor dataset based upon policy-relevant criteria common to both datasets (e.g., dwelling tenure, employment status, income class). The information on donor records thus classified may then be attributed to records with similar characteristics on the host dataset without the possibility of adding to their identifiability.

Figure 1 provides an overview of the SPSD creation process. The ellipses represent data files (e.g., the SCF, the Green Book) and the rectangles represent processes. We turn next to the main part of the paper where each step in the construction of the SPSD, as shown in Figure 1, is described.

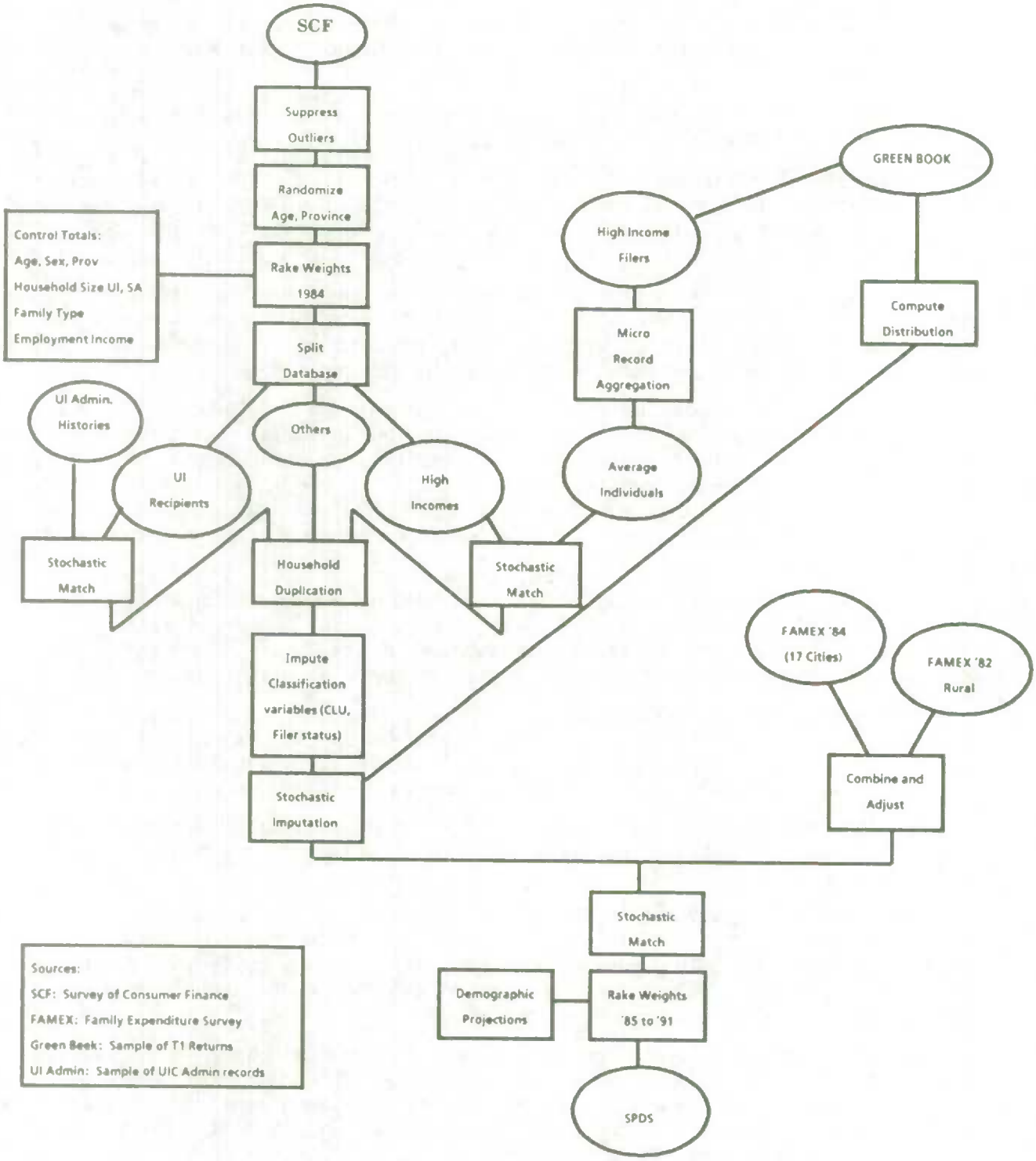


Figure 1: SPDS Database Creation Process

3. THE HOST DATA

The target or "host" dataset is derived from the 1984 Statistics Canada ***Survey of Consumer Finances*** (SCF), an annual survey administered to selected households drawn from the survey frame of the Labour Force Survey (LFS). Four different forms are collected from each sampled household. The Household Record Docket contains demographic information on each individual in the household, as well as family structure information. The LFS form contains information on the labour force status for individuals aged 15 and over in the household. The SCF form has the income, by source, for each member of the household aged 15 and over. The Household Income Facilities and Equipment (HIFE) form details the characteristics of the dwelling, and certain kinds of equipment contained in it. In 1984 the survey consisted of approximately 36,000 households containing 98,000 individuals.

Associated with each household in the sample is a Record Docket and a HIFE form, and associated with each individual in the household aged 15 and over is an LFS form and an SCF form. Because of the great wealth of already linked information that results, this combined hierarchical database forms the starting point for the SPSD creation process.

It may be noted that even though these diverse data are fully integrated at the microdata level, in the early production phases of the survey, the public so far has never had access to this rich multivariate information. The survey results emanate from Statistics Canada as distinct public use sample tapes or print publications on individual incomes, economic family incomes, census family incomes, HIFE and the labour force survey. This traditional and fragmented view of the utility of microdata sets is one that is being challenged by the SPSD. Our objective is a fully hierarchical database including individuals, census families, economic families and households.

The information from the UI, Greenbook and FAMEX files was then "added" to the SCF. In order to exploit fully this information being imputed from other sources, many original SCF records were cloned or duplicated. For example, records representing unemployed individuals were duplicated until the number conformed to the sample size of the UI file (about 30,000). Records representing high income individuals (those with an income of over $80,000 in 1984) were duplicated to correspond to the number of high income records derived via micro-record aggregation from the Revenue Canada sample (about 5,000). To maintain the family structure and overall sum of weights, the records of all other persons in households containing either unemployed or high income individuals were similarly duplicated. The weight assigned to a record was reduced to account for the number of times it was duplicated. The resulting database contains over 170,000 records with a high proportion of the records representing households containing unemployed or high income individuals.

3.1 Suppression of Outliers

A guarantee of the non-confidentiality of the constructed database (SPSD) is provided if each input microdataset is itself non-confidential, and if data "merging" does not involve exact matching. This is the strategy that has been adopted, and begins with screening the SCF file.

Public release versions of the host (SCF) data are pre- screened for potentially sensitive cases. For example, households with more than nine members are deleted from the public release household file, and census families with more than four UI recipients or more than 6 earners are deleted from the public release census family file. The initial step in SPSD database construction was to suppress each household that met any of the SCF screening criteria (i.e. the criteria applied at the household, economic family, or census family levels).

In addition to suppression of entire households, certain SCF recodes were performed. These involved, for example, merging certain geographic areas (e.g., Brandon with Winnipeg) or recoding as unknown the occupation codes for spouses of high income individuals.

3.2 Randomization

Further protection against release of identifiable households is provided by age-sex and regional randomization.

It is assumed that disclosure of the age-sex composition and location of a household may increase the risk of a breach of confidentiality. However, this risk may be considerably reduced by randomizing the ages of household members within five year age groups and by randomizing the sex of children (i.e. aged <= 15).

Similarly, the location of unusual household types may be changed by randomly reassigning their province and urban size class codes. Unusual household types are equated with households containing more than eight individuals, more than 2 census families, more than one economic family, or individuals with special income or tax characteristics (e.g. females with income above $80,000, or male or female with income below $150,000 and income tax greater than $150,000).

3.3 Iterative Proportionate Adjustment (IPA)

Given that the SPSD database includes complete household and family structures, it is essential to associate a single weight with each household that will guarantee consistency in tabulations at the household, family and individual levels. This is not done at present because Statistics Canada public release databases are provided with separate weights at individual, census family, or economic family levels.

In order to provide this consistency, multi-level IPA was employed. The procedure is a generalization of the ordinary IPA (popularly termed 'raking') procedure employed on the SCF to obtain individual level weights that are consistent with known age-sex control totals. It may be thought of in terms of successive (proportional) adjustments to survey weights to bring them in line with pre-determined control totals. In multi-level IPA, the adjustments may be applied at household, family, and/or individual levels with an additional step which replaces individual (adjusted) weights within a household by the household average.

In addition the SCF also exhibits reporting biases which restrict its utility for modeling tax and transfer programs, for example:

- non-reporting of high-income individuals,
- under-reporting of social assistance income, and
- under-reporting of investment income.

Using iterative proportional adjustment, the SCF record weights were recalculated to correspond to external control totals such as the number of high income (over $80,000 in total income in 1984) individuals, family size by province, private pension income and Social Assistance benefits by province.

The control totals employed in constructing weights for SPSD represented: (a) individuals by age and sex, (b) individuals by income class, (c) individual UI claimants, (d) households by family composition and labour force participation, (e) households by Social Assistance benefits, and (f) individual pensioners. Each of these control totals was disaggregated by province.

It has been shown (unpublished research by Georges Lemaitre, Social Survey Methods Division, Statistics Canada) that IPA adjustments of this sort lead to improved estimates of population characteristics. In particular, use of multi-level IPA with control totals provided only by population by age and sex produces estimates of family level characteristics with a 50% reduction in sampling variance compared with the principal person method currently employed. While, at the individual level the sampling variance is essentially the same as current methods produce.

3.4 Splitting Database

Splitting refers to a mechanical data preparation step that partitions the SCF (after suppression of outliers, randomization and IPA) into three mutually exclusive subsets: high income individuals, UI recipients, and all others. To simplify subsequent steps in the database creation p this split is done in such a way that no households containing high income individuals also contain UI recipients. There are, in fact, a handful of such cases but UI recipients in these households are treated as though they received no UI. High income individuals are those with incomes over $80,000 while UI recipients are those who reported receiving some benefit in the SCF survey.

4. STOCHASTIC MATCHING

Stochastic matching involves creating 'fused' composite records from two micro-data databases. Consider two databases, a host database A and a donor database B. There are a variety of methods that can be used to attribute some or all of the information on a record from database B onto any given record from database A. All are based on the idea that we wish to find a record from database B which is in some sense similar to the given record from database A. The determination of similarity is based upon variables common to both databases and is affected by the intended use of the 'fused' records. Various 'nearest-neighbour' algorithms, which use methods similar to those of cluster analysis, can be used to determine a mathematically 'optimal' match, given a particular method of determining distance in N-dimensional space. Complications arise in practice due to limitations on the size of the set of 'donor' records (database B in our example) and the desire to use non-continuous variables (e.g. discrete or categorical).

In the SPSD a different more heuristic technique was used. It involves partitioning the two databases into identically-defined 'bins' of records, which are then sorted based upon one of the continuous variables common to the two databases (usually total income in SPSD). Records in a given bin are then matched one-for-one across the two databases (i.e. record n in bin m of database A is matched with record n of bin m in database B). Complications arise because the number of records in a given bin is generally not equal in the two databases, and also as a result of the presence of record weights on one or both databases. These problems are solved by selectively duplicating records from one or both databases.

The SPSD uses stochastic matching for adding FAMEX data, UI data, and Green Book income data for high-income recipients. The technique allows the preservation of inter-item correlations from the donor record. Each of the matching procedures is described more fully below, where it is also noted that these stochastic matches virtually preclude the possibility of an exact match.

5. HIGH INCOME ADJUSTMENT

The SCF has known reporting and sampling biases which result in a lower number of high-income individuals and fewer dollars of income per high-income individual then implied by personal income tax records. In the creation of the SPSD, both under-reporting and non-reporting of several income and deduction items are dealt with. Figure 2 provides an overview of this high income adjustment process.

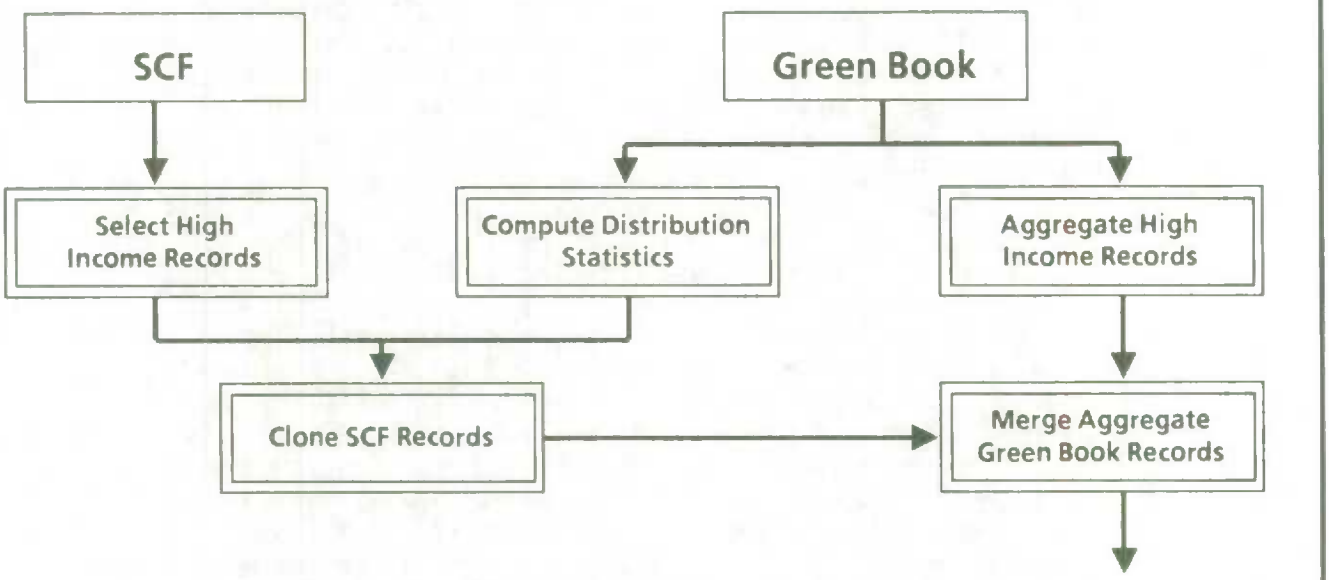


Figure 2: High Income Adjustment Process

5.1 Micro-Record Aggregation

Non-reporting by high-income individuals in the SCF is ameliorated by using the Green Book counts for individuals with income over $80,000 as a IPA margin. The IPA then increases the weights of each high-income record on the SCF so that the sum of the weights corresponds to the Green Book.

There are approximately 300 such records. The IPA process leaves them with very high weights (on the order of 200-500). These records are used as the "hosts" for accepting the more precise information from the Green Book. This in turn provides the basis for an adjustment of income items for the high-income group.

Even with a scaling up of the weights for high income records on the SCF, there is still a substantial under-reporting of income in this group. As a second step, under-reporting bias is corrected by replacing the income components on these records with plausible but non-identifiable sets of income items from the Green Book (see Table 1).

Table 1

SCF Income Items Replaced for High Income Individuals

Employment Related
Earnings from Employment
Farming Net Income
Other Allowable Employment Expenses
Self-employed Income - Non-farming
Investment Related
Allowable Other Years Capital Loss
Allowable Prior Years Non-capital Loss
Carrying Charges
Capital Loss on Disposition of CCPC Equities
Interest Income
Net Rental Income
Other Investment Income
Taxable Capital Gain/Loss For Year
Taxable Amount of Canadian Dividends
Other
Other Taxable Income
Imputed Total Income - Sum of Components

Records from the Green Book are grouped into sets of at least 5 records. These grouped records are considered to be a non-confidential table although they retain many of the characteristics of micro records. The groups represent individuals of similar age, employment income, investment income, dividend income and capital gains. For these groups, or five-tuples, a weighted average is calculated for the items listed in Table 1. Once grouped, the records are considered non-confidential since they represent 5 or more individuals. This is equivalent to publishing a table in which each cell contains no less than 5 individuals.

The resultant aggregate contains 4,676 records representing 24,556 Green Book Records, in turn representing 133,650 high-income filers. These aggregate records, derived from otherwise confidential microdata, are now able to become part of a public use data set with little loss of information.

5.2 Stochastic Match

The original 300 SPSD records are duplicated to match the number of aggregated Green Book high income records (4,676). These 300 records do not provide a sufficient basis for the demographic characteristics of the high income filer population. Thus a detailed match by age, sex, province and total income would not be feasible. Instead, the duplicated SPSD records were imputed a new value of total income based on a very simple age break (2 groups), sex and region using the same procedure described in a subsequent section (Stochastic Imputation of Income Tax Information). This new imputed value of total income was used as a key to sort the SPSD records before merging the similarly sorted, aggregate Green Book records.

To improve the match with regard to age, sex, province, total income and tax status, a much larger original SCF sample would be required.

5.3 Evaluation

Although this method of micro-record aggregation assures that correlations between the income and deduction items (shown in Table 1) are generally maintained, the univariate distributions of the synthetic records tend to have less variance than the original Green Book records. This is a result of the aggregation of several records into one. Very often, for sparse items such as Allowable Other Years Capital Losses, the five records to be aggregated contain several zeros which are included in the average. The average is maintained but the distribution is biased towards the mean.

Figure 3 provides an example of the distortion in the distribution of Capital Gains introduced by this method. In effect, five-tuples of individuals were all attributed small values for Capital Gains instead of four with zero values and one with a higher value.

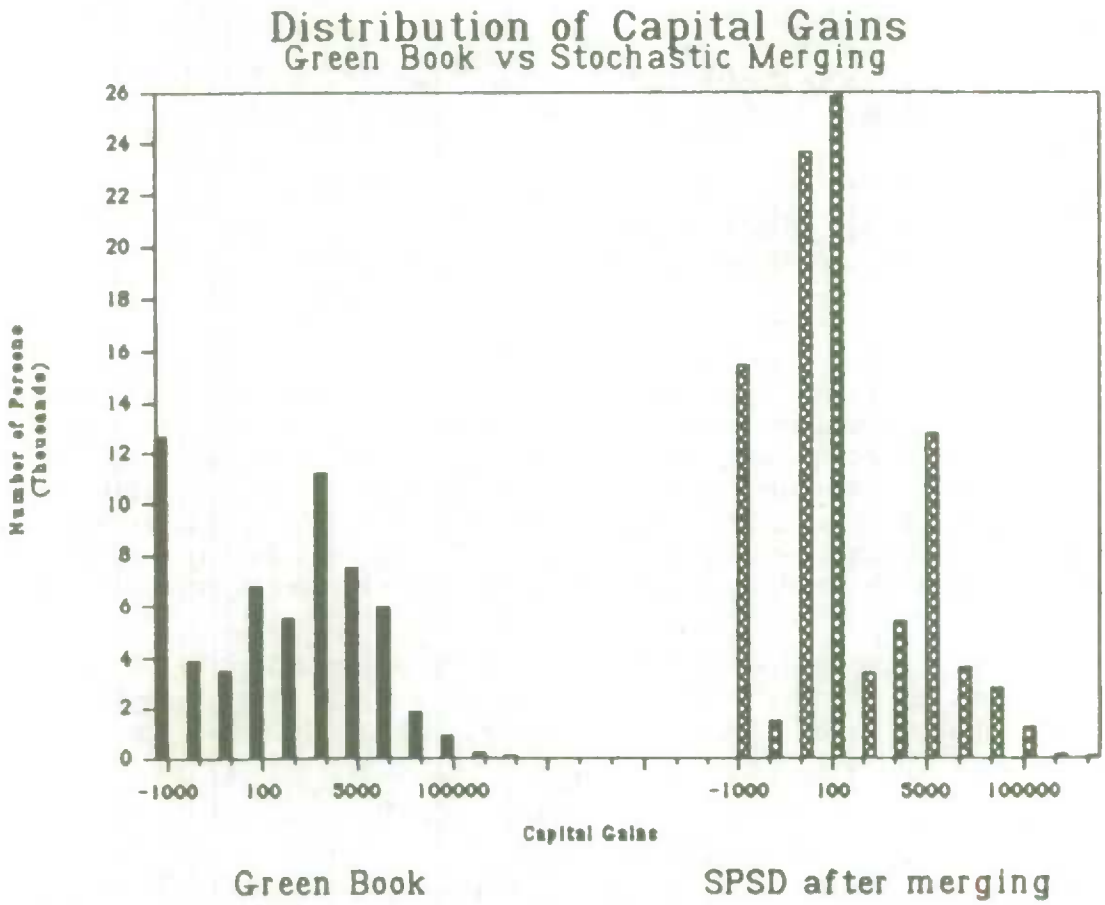


Figure 3: Capital Gains Pre and Post Duplication

6. UNEMPLOYMENT INSURANCE HISTORY IMPUTATION

Unemployment Insurance (UI) is a complex insurance and temporary income maintenance program, the administration of which requires monitoring claimants' weekly labour market activities. The administrative data collected under the program serves (i) to track the weekly benefits and claim activity of UI recipients, (ii) to establish eligibility and entitlements by monitoring previous program participation in the event of repeat or re-entrant claims and (iii) to monitor past employment patterns through "Records of Employment".

The UI model was intended both to provide a stand alone model of a major transfer payment program, and to provide a simulated UI income component as input to the taxation model. As a stand alone model, the required output serves to indicate program costs, client population, gainers and losers under alternative program structures, and to permit examination of financing issues. Input to a taxation model is provided by simulated benefit payments on a calendar year rather than a claim basis. Thus, the initial task in constructing this component of the database required simultaneous development of a UI simulation module and identification of a limited set of UI variables (Table 2) that could serve as input to the module.

6.1 UI Donor Dataset

The UI administrative histories imputed to SPSD were based on a 1% sample from the population with some claim activity within the 1984 calendar year. The sample consists of about 30,000 individuals and represents about 40,000 claims. The content of this dataset was specially designed. On one hand, it had to be rich enough to capture the weekly labour force history relevant to application of UI program regulations. On the other hand, it had to be compact and general enough to be non-confidential. This was accomplished by thinking in terms of an event history, so that the durations of various activities became the focus rather than weekly activity records. The staffs of Employment and Immigration Canada and of the Forget Royal Commission were helpful in designing this dataset. Table 2 shows the set of variables employed as input to the UI model.

Table 2

UI History Variables

Claim Sequence Number (1st or 2nd current year)
Repeater Flag
Initial Benefit Type Type Change Flag
Weeks of Benefits (current claim)
Weeks of Benefits (in previous 52 weeks)
Weeks of Work (prior to current claim)
Average Weekly Earnings (prior to claim)
Penalty for Voluntary Quit (weeks)
Week Claim Established
Benefits Paid in Calendar Year (1 or 2 claims)
Weeks of Benefits Paid in Calendar Year

Because of the interrelatedness of these UI claim history variables, each of the 30,000 claimants' records (which may consist of one or two claims) was stochastically matched to SCF records which had some reported UI income in the calendar year. In addition to the

UI claim history variables identified in Table 2, administrative data on claimant age, province and sex are used as matching keys. These same variables were available on the host dataset for individuals with UI income.

Claim types are an important element in the match, since there are currently major differences in eligibility rules and in entitlements between types. A claim type classification was constructed on the host dataset by (i) identifying UI recipients aged 65+ (Retirement benefits), (ii) identifying UI recipients with occupation coded as "Hunting, Fishing, Trapping" (Fishing benefits), and (iii) identifying female UI recipients with a child aged 0-1 (Maternity benefits). No distinction could be made between Sickness and Regular benefit types on the host dataset.

6.2 Stochastic Match

Matching was carried out by first partitioning the donor administrative (UI) and host (SCF) datasets on the basis of age group, province, sex, and claim type. Duplication of records within cells was carried out to ensure that corresponding cells of the UI and host datasets had equal numbers of records. If in any given cell the number of host records exceeded the UI records, then the UI records were uniformly duplicated (UI data were a simple random sample). Correspondingly, if the number of UI records exceeded host records, then host records were duplicated in proportion to their weights (recall that the host data were based on a stratified sample). The latter case was the more frequent condition (in 170 out of 218 cells), but the former also occurred (a consequence of stratified survey design). Duplicates of host dataset records had weights adjusted in proportion to the number of times that they had been duplicated.

The outcome of the cell match and duplication steps was an increase in the number of records representing the UI claimant population. Initially, the host dataset contained 10,381 such records, while after duplication there were 31,585 records. This expansion of the dataset was intended to ensure full use of the UI histories available from the 1% sample.

Within cells, matching host and UI records were identified as the records with corresponding rank in the two datasets. The records were ranked on the UI benefits received (in dollars).

6.3 Evaluation

Table 3 provides an indication of the success of the match. The correlation between benefits reported on the SCF and the corresponding (matched) benefits from the donor UI dataset indicates the 'accuracy' of the match, since benefit ranks rather than benefits per se were used in the match. Difference quartiles represent the 25%, 50% and 75% cutpoints for the distribution of differences between SCF and UI benefits.

In most cases, the differences between benefits reported on the host dataset and benefits sampled from UI administrative data are relatively small. Discrepancies as large as 255 dollars may be expected, since they could represent a UI benefit payment for a single week (i.e. the minimum discrepancy in benefit weeks). Moreover, the differences are small in comparison to median benefit levels, which were $2,972 for males and $2,050 for females, at the national level.

It is expected that some host data may represent biased responses and that others may contain benefit components not included in the UI data or model (e.g. training allowances). If this were the case, then error in the host dataset would make an important contribution to the benefits differences.

Correlations are high, except in PEI where little gain from duplication was possible. In the absence of substantial duplication, age and claim type matching constraints will reduce marginal correlations in benefits.

High correlation is not a necessary consequence of the matching technique. Matching of corresponding ranks guarantees a monotone association, but not necessarily a strong linear association. This use of ranks can be interpreted as matching corresponding quantiles of independent samples. Thus, a strong linear association indicates that the two samples (host and donor) are from similarly shaped density functions (i.e. belong to the same location-scale family).

Further direct evaluation of the results of the match is difficult, since essentially all common factors between datasets have been employed in the match. The UI data provide an extension and replacement of host data in which UI variables are unbiased and consistent with the UI program structure.

Table 3

Comparisons Between Matched UI Records

Distributions by Province & Sex and for Canada
(i) n - Number of Pre-duplication Records
(ii) r - Correlation Between Host & Donor UI Benefits ($)
(iii) Difference Quartiles - [Host($) - Donor($)]

Province/Sex		n		r	Difference Quartiles		
		Host	UI		25%	50%	75%
NFLD -	Male	795	929	0.953	-192	140	417
	Female	445	549	0.925	-270	-14	232
PEI -	Male	241	246	0.631	-1,159	-290	789
	Female	210	213	0.871	-363	11	531
NS -	Male	496	787	0.931	-271	45	528
	Female	294	528	0.919	-197	-38	147
NB -	Male	604	798	0.941	-531	-45	589
	Female	390	573	0.905	-102	158	669
QUE -	Male	1,116	5,471	0.970	-162	86	341
	Female	784	3,961	0.958	-112	103	324
ONT -	Male	787	4,990	0.960	-149	36	207
	Female	687	3,837	0.953	-110	74	306
MAN -	Male	343	611	0.932	-360	-69	294
	Female	272	508	0.866	-115	-49	496
SASK -	Male	369	548	0.918	-239	231	489
	Female	283	394	0.954	-83	75	311
ALTA -	Male	691	1,648	0.946	-88	68	448
	Female	482	1,072	0.951	-174	16	264
BC -	Male	625	2,281	0.953	-112	186	470
	Female	467	1,638	0.954	-185	68	461
CANADA		10,381	31,582	0.953	-155	69	352

7. HOUSEHOLD DUPLICATION

There are three conditions under which duplicates of SCF household records are created. These are: (1) in the imputation of taxation data to high income earners, (2) in the stochastic matching of UI data, and (3) in the creation of synthetic Institutionalized Elderly.

In the case of taxation or UI data, the motivation for household duplication is to utilize as much of the variety in the administrative data as is possible. Note that in both of these cases, duplicates of individuals are formed first. Then the other individuals in their household are also duplicated. In the event that more than one member of the same household is duplicated, then additional duplication is necessary to ensure that each individual is properly represented.

Institutionalized Elderly are created by direct duplication of non-institutional unattached elderly (aged 65+) who are not labour force participants. The motivation for selecting this donor population is that these individuals may be most at risk of institutionalization, and thus their income characteristics are most likely to resemble the institutional population. The weights on these records are adjusted to reflect estimates of the institutional population by age, sex and province (based on administrative statistics on institutional bed days).

8. STOCHASTIC IMPUTATION OF INCOME TAX INFORMATION

This section will describe stochastic imputation, the method used to attribute personal income tax information to the SPSD records. The following list of deductions and income items were imputed from the Green Book onto the SPSD. These are items which are not well represented on the SCF (e.g., capital gains), entirely absent (such as carrying charges) or not easily modeled (e.g., disability deduction).

1. Other Allowable Employment Expenses
2. Carrying Charges
3. Child Care Expenses Allowable
4. Charitable Donations and Gifts
5. Allowable Other Years Capital Loss
6. Disability Deduction
7. Union and Professional Dues
8. Education Deduction for Student
9. Other Federal Tax Credits
10. Federal Political Contribution Tax Credit
11. Taxable Capital Gains
12. Capital Loss on Disposition of CCPC Equities
13. Federal Investment Tax Credit
14. Net Medical Calculated Amount
15. Allowable Prior Years' Non-capital Loss
16. Other Deductions from Net Income
17. Other Dependent Exemptions

18. Provincial Tax Credits
19. Total RPP + RRSP Contributions
20. Proportion of RRSPs in (RRSP + RPP)
21. Tuition Fees

These items, in combination with other provisions which can be readily computed from available data (e.g., personal exemptions) allow a complete calculation of taxable income and tax payable.

8.1 The Source Data

The source data for the imputation were derived from a Revenue Canada sample of 1984 Individual Tax Returns. This contained 2.4 percent of all returns (380,419 returns), the same sample used to compile the ***Taxation Statistics*** (the Green Book) publication. The sample is stratified by source of income, urban geographic area, rural geographic area, tax status (taxable and non-taxable), and income range.

The information in this sample contains most of the information submitted in the 1984 T1 Federal and Provincial Individual Income Tax Return and accompanying schedules. This sample has no explicit family structure (i.e., the returns of the head, spouse and dependents cannot be analyzed together in an identifiable family unit).

8.2 Data Transformations

To join these Green Book income tax data with the SCF-based host sample a set of common classification characteristics were defined. The following attributes were chosen as much for their degree of policy relevance as for their availability and similarity of definition on both datasets:

1. Taxing province
2. Age group
3. Sex
4. Marital status as taxed
5. Total Income class (excluding Capital Gains)
6. Employment Income class
7. Children claimed for the Child Care Expense Deduction (on SCF, number of children eligible for claiming).

Sub-samples defined by the cross-classification of these items are assumed to have sufficiently different distributions to merit retaining the uniqueness of these distributions. Figure 4 provides an example of the difference in capital gains between two income groups. A comparison of charitable donations between the same groups is provided in Figure 5.

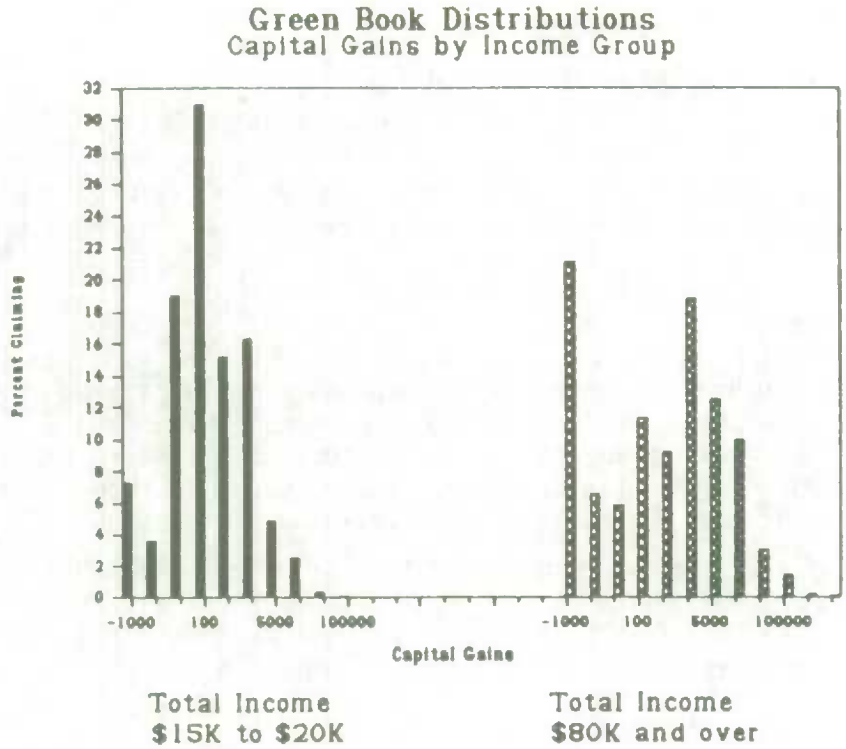


Figure 4. Green Book Distribution of Capital Gains for Two Income Groups

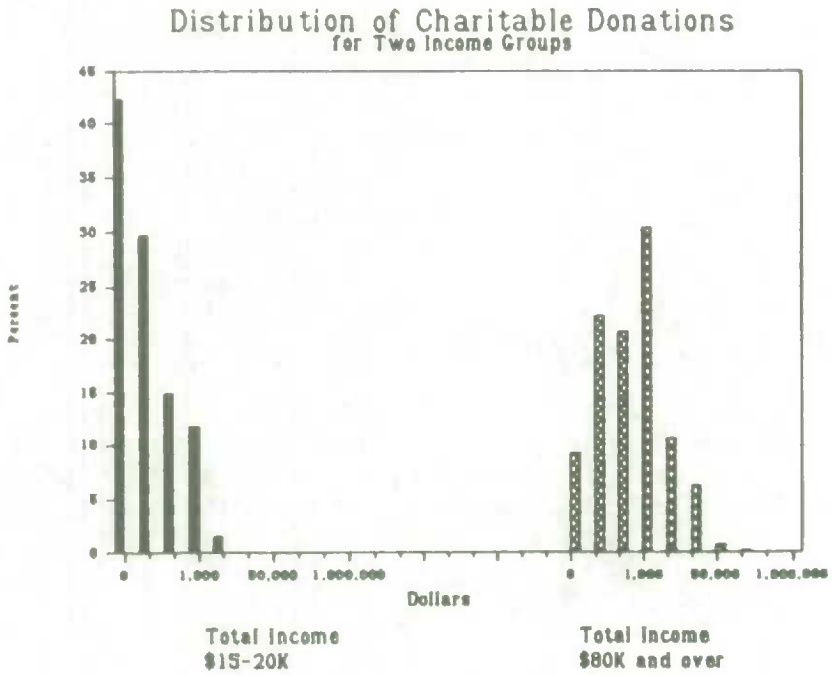


Figure 5. Green Book Distribution of Charitable Donations

Prior to imputation, the host dataset was prepared by identifying potential tax filers, establishing eligibility for certain targeted items (Education, Tuition and Child Care Expense Deductions), and creating a parallel classification scheme on both the host SPSD and donor Green Book datasets.

A model of the personal income tax system (the same one subsequently used for policy analysis) was initially employed to identify likely tax filers and to impute marital status as taxed. For example, a married person eligible to claim his or her spouse as a dependent, would be designated married-taxed-married. This imputation was essential to restrict the imputation to a similar universe as the donor dataset.

Three of the deduction items were treated specially in that the eligibility for these items could be identified on the host dataset. From information available on the SCF, one is able to determine if the individual is eligible for the Education Deduction (self or dependent is attending a post-secondary educational institution), Tuition Deduction (self is attending a post-secondary institution) and the Child Care Expense Deduction (for lower income spouse with children under 15 present). Targeting the imputation to individuals eligible for these deductions ensures some degree of internal consistency in the synthetic records. For example, only persons with children will be imputed the Child Care Expense deduction.

Unfortunately it is not as simple to determine eligibility for all deductions and income items imputed.

The joint distribution of RPP (Registered Pension Plan) and RRSP (Registered Retirement Savings Plan) contributions posed a problem in that the tax law restricts the total of the two to be below a certain limit ($3,500 in 1984). Imputing the two separately would not ensure that this threshold is not exceeded. To overcome this, we imputed the sum of the filer's RPP and RRSP contributions, and then RRSP contributions alone as a proportion of this sum.

8.3 Deriving Distributional Statistics

One purpose of this imputation process is to duplicate the distribution of deductions and income items as found in the Green Book file. This requires a method of representing an arbitrary distribution. For example, the method should equally well represent bimodal, truncated and long-tailed distributions.

Another factor in the choice of method was its computational intensity. Since the source dataset contains almost 400,000 records, the algorithms to generate these statistics had to be reasonably efficient.

The method eventually chosen was to represent the univariate distributions of particular items first by the proportion in any given sub-group with a non-zero value for the item, and then for the sub-sub-group with non-zero values. The density function was represented by their decile cut-off points with special treatment of the tails of the distributions.

The same procedures were applied independently for two sets of statistics: percentage reporting and distributions. They were treated separately because the percentage reporting required a less stringent rejection criteria and therefore information from a lower level of aggregation could be used.

The percentage reporting statistic was kept if the sum of weights for the cell exceeded 400 or the number of records representing a non-zero value exceeded 20. If these criteria were not met, the statistics for a higher level of aggregation was substituted.

The criteria for the distribution statistics had to be more rigorous. The minimum cell size for was 100 records, i.e., if a cell did not contain at least 100 non-zero records, statistics for that cell were not computed.

For each item to be imputed, the nearly 400,000 income tax return records were classified into relevant cells (e.g., income group by age by marital status by sex by province).

For each cell, given a sufficient sample, the following statistics were computed:

- values for decile cut-points 1 through 9,
- the mean of the bottom and top deciles,
- the mean of the highest 5 values and the mean of the lowest 5 values, and
- the percentage within the cell reporting a non-zero value for the item.

These statistics are well suited for representing an arbitrary distribution and they are simple to calculate.

For confidentiality reasons, the actual maximum and minimum values in a cell could not be used. The mean of the highest five values and the mean of the lowest five values in the cell were used as substitutes.

The same statistics were then generated for aggregations of cells, in this case, for income group by age by marital status by sex by region. Collapsing the 10 provinces into 5 regions increases the level of aggregation and therefore increases the average number of individuals within a cell. More cells will then provide valid sets of distributional statistics.

Ideally, all values would be imputed from the lowest level of aggregation. However, due to the sparseness of many of the data items this is rarely possible. For example, Other Allowable Employment Expenses are concentrated in the higher income groups and cells in this region would be well represented. For the lower income groups, the cells are sparser and often empty.

To fill in these sparse and empty cells, statistics from higher levels of aggregation are substituted. If, for instance, the cell representing the following classification:

-Income Group	$35,000 to $39,999
-Age Group	25 to 35
-Marital Status	Single, Taxed Married
-Sex	Female
-Province	Quebec

were empty or rejected on the size criterion, statistics would be substituted from the next level of aggregation:

-Income Group	$35,000 to $39,999
-Age Group	25 to 35
-Marital Status	Single, Taxed Married
-Sex	Female

representing this income group, age group, marital status and sex for all of Canada. If this cell were also sparse or empty, statistics would be substituted from the next higher level of aggregation. In the worst case, the statistics for a cell would be derived from the entire sample, i.e., all income groups, all age groups, all marital statuses, both sexes and all provinces.

The resultant distribution and percentage reporting statistics are non-confidential since they never reveal raw data values. The extreme values are synthesized by calculating the mean of the highest 5 values and the mean of the lowest five values.

8.4 Imputation

Using this complex set of distributional statistics generated from the Green Book, it is possible to recreate the same distribution of values on the host dataset. For each eligible individual on the host dataset, a synthetic value is drawn from a distribution representing the tax returns of a similar group of people.

Values for the middle eight deciles are generated assuming a uniform distribution between decile cut-off points.

The top and bottom deciles are treated specially so that both the shape and the size of the tails are accurately represented. Preservation of the tail of the distribution is essential to maintaining overall means and totals, especially for items with long-tailed distributions such as capital gains or business losses.

In imputing the upper and lower decile, values are drawn assuming a Pareto distribution to generate the appropriately shaped tail. The mean of the top decile is maintained, thereby preserving the total size of the tail. Extreme values are truncated at the mean of the highest 5 values in the group. The same procedure was applied to the lower tail of the distributions.

8.5 Evaluation

Figures 6 and 7 provide some examples of results of the imputation process. These are both aggregated to the level of the entire sample.

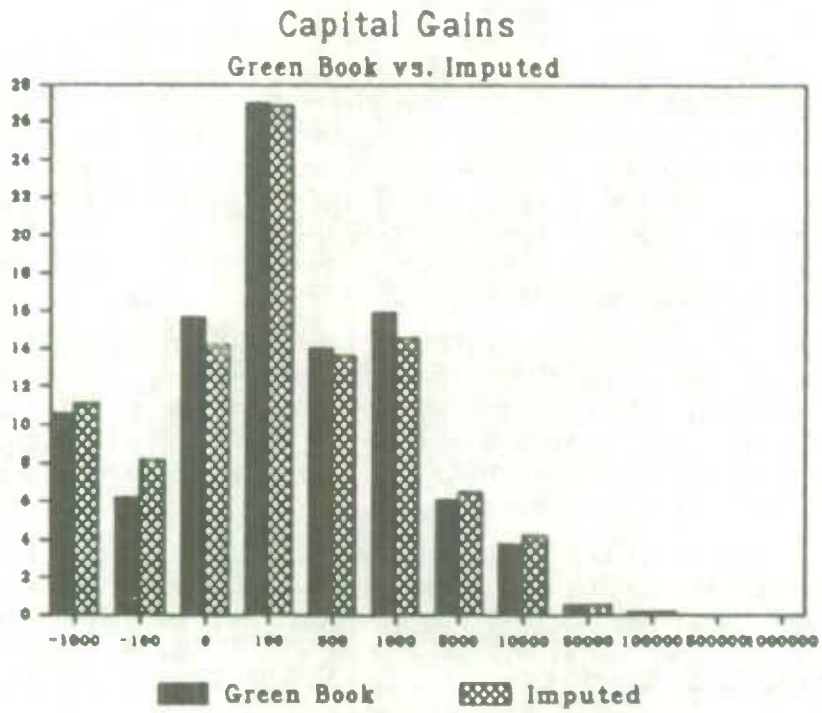


Figure 6: Capital Gains Distributions Pre/Post Imputation

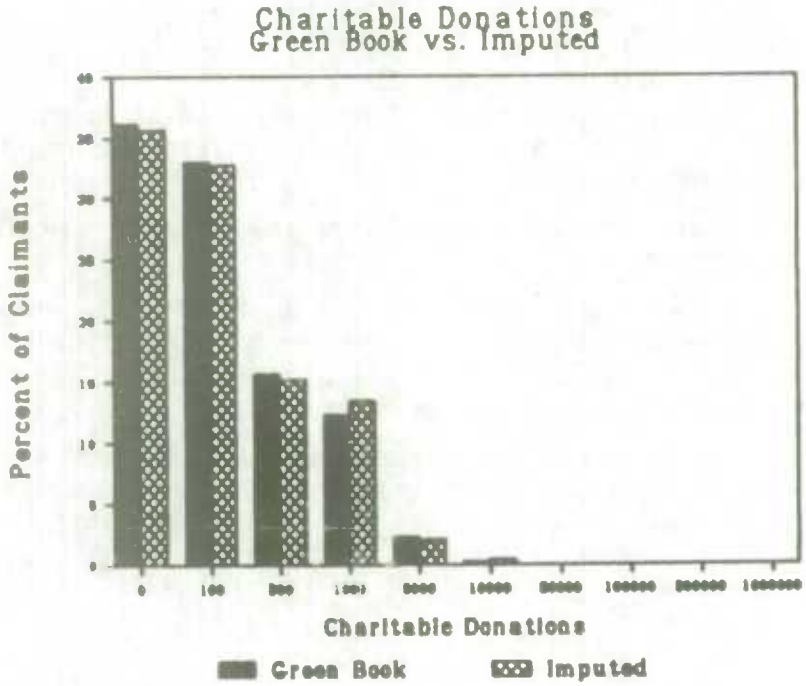


Figure 7: Charitable Donation Distributions

This method of imputation tends to use the full richness of the source data in regenerating plausible distributions on the host dataset. Overall distributions make sense but often individual cases do not. For example, since Capital Gains are imputed according to total income, age, sex and province, it is not impossible for a Social Assistance recipient to be imputed Capital Gains. In this case, the Social Assistance recipient is treated exactly the same as retiring farmer who has sold his farm and has received several hundreds of thousands of dollars in Capital Gains.

Another problem with this method is that joint distributions are lost to the degree that they are not accounted for by the classification variables. In simple cases, most deduction items are well correlated with income and income is normally an important classification criterion. Where the inter-correlation between items (e.g., RPP and RRSP Contributions) is more important than their correlation with income, the correlations are lost unless the method is modified.

One outcome of the loss of correlation between deduction items is that the individuals, especially the high income group, on the host dataset appear not to be optimizing their tax situation. Since the high income group consists of all individuals with income over $80,000, an individual with a total income of $90,000 has the same probability of being imputed a million dollar deduction as another person with $2 million in income.

In the next version of the SPSD, this problem may be overcome by applying the micro record aggregation method to impute more deduction items for the high income group. This would have the effect of preserving correlations between income and deductions as well as correlations among deductions.

This would have the effect of preserving correlations between income and deductions as well as correlations among deductions.

9. FAMILY EXPENDITURE SURVEY DATA IMPUTATIONS

The family expenditure data are intended to support simulations based on shelter costs (e.g. Social Assistance), simulations concerned with child care costs, and simulations of commodity taxes. Due to the limited number of records on the family expenditure dataset it was decided to perform three separate imputations to allow for a specific tailoring of the classification categories to the nature of the vector of expenditure items to be matched. For example, a household's expenditures on child care depends substantially on the number of children and the labour force status of the parents, and as such these should be the primary classification variables in any match. On the other hand shelter costs are more strongly correlated with the number of rooms and residential tenure and a classification by number of children would do little to improve the match.

Four main steps were involved for each of the three imputations.

- Construction of a National 1984 FAMEX database
- Selection/Grouping of expenditure items for imputation
- Selection/Construction of Matching Variables
- Stochastic Matching (Weighted Duplication)

9.1 Inflating the 1982 FAMEX

The family expenditure survey was last conducted for all of Canada in 1982. The 1984 survey, which matches the time frame of the SCF host dataset, was restricted to a 17 city sample. The first step in matching was to prepare a 1984 all Canada version by inflating the values of the 1982 non-17 City records to 1984. All money items on a given 1982 family expenditure record were "grown" simply by using the same inflator rather than using individual CPI and income inflators. This simple process was dictated by the requirements of the commodity tax model where a complete accounting identity of a household's income, expenditure and saving patterns must be maintained.

This approach assumes that expenditure patterns of households remain constant and thus avoids implicit assumptions of behavioral response to price fluctuations. This assumption was supported by an analysis of shifts in the proportions of total expenditures spent on individual items between the 1982 and 1984 17-City samples. The differences between all items remained within one percent. The largest differences were a one percent increase in mortgage interest as a percent of total expenditures (4.7 to 5.7), and a .6 percent decrease in automobile and truck purchases (5.4 to 4.8).

The inflators were computed individually for each non 17-city record on the 1982 FAMEX. They were based on average growth by family type in each of six income sources as reported on the Survey of Consumer Finances. A household's inflator was calculated as an weighted average of the six individual growth rates for their family type where the weight was the proportion of income received from each source.

9.2 Determination of Imputation and Matching Variables

Table 4 summarizes the variables for imputation as well as the matching variables used in each of the three stochastic matches. The figures in parentheses represent the number of classification levels.

Table 4

Variables and Classifications for FAMEX Match

	Shelter(126)	Child Care(36)	Expenditure Vector(390)
Imputed Variables	Rent Mortgage Interest Property Taxes Insurance Premiums Utilities Repairs Other Shelter Costs Value of Home Balance on Mortgage	Child Care Expenses	"Savings" Other Money Receipts Household Income Account Balancing Difference Expenditure Vector (50) (See Appendix A)
Matching Variables	Residential Tenure Number of Rooms Urbanization Geographic Region Household Income	Family Type Employment Status # Children (0-4) # Children (5-15) Household Income	Income (discrete 6) Family Type(5) Residential Tenure(3) Age of Head(4) Sex of Head(2) Geographic Region(5) Family Size(2) Number of Children(3) Urbanization(2) Income (continuous)

The variables selected for the shelter match were selected and grouped so that estimates of major costs and imputed rent could be made. The chief intended use was for modeling social assistance payments and secondarily for use in modeling provincial tax credits. The high level of aggregation reflects the coarse way in which social assistance can be modeled due to the lack of other data relating to eligibility and benefit levels. For example asset eligibility tests are made for a five year period and benefit levels are based on fire insurance while FAMEX reports only total home insurance.

Child care costs are composed of day care costs inside or outside the home as well as kindergarten tuition fees. This definition is intended to follow current federal legislation regarding the child care expense deduction. No attempt has been made to exclude costs that may be disallowed for tax purposes due to the absence of receipts. Other items such as infant's clothing or other variables which may be desired when modeling an expanded definition of costs are not imputed.

The selection and grouping of FAMEX income and expenditure variables for the expenditure vector was based on the structure and composition of the personal expenditure dimension of Canadian input-output tables and the requirements of the commodity tax model. Expenditures having some indirect taxes and duties were placed in the corresponding input-output personal expenditure category. Variables not having an indirect tax, or an indeterminate indirect tax were placed in a residual category (e.g. real estate commissions). Additional variables were also included in the vector (e.g. income, taxes, savings) in order to complete an identity to allow for various simulation options. Conceptual differences between FAMEX and the system of national accounts on which the input-output tables are based have not been adjusted for during the match.

The determination of matching variables was restricted by the availability of similar variables on both the host and donor datasets. From this limited set, individual analyses were conducted to determine the optimal selection and configuration of the matching variables for the three matches. The techniques used to identify variables included correlation, factor analysis and difference of means tests. Four main interdependent criteria guided the selection and creation of matching bins:

Expenditure Levels: The variables used for classifying households should be highly related to both the level of expenditure as well as the distribution among specific commodity elements.

Expenditure Categories: The bins should be created in such a way as to allow for the attribution of costs to only appropriate populations. For example childless couples should not have child care expenses and unattached women should not have large men's clothing expenses.

Reporting Categories: The bins should reflect to as great a degree as possible the categories that will be used in final reporting. For example the SPSD and model are likely to be used for comparative analysis of different provinces and regions, different levels of income, and different family types, so these variables should be used in the matching process.

Sample Size Within Bins: when creating the bins it was an essential requirement that both the host and donor databases had at least five observations in any bin. In practice some bins contained very large numbers so that the final sort on income was a key element of the fit for all three matches.

The analysis and criteria were taken into account in creating the final selection, groupings, and priorizations. The hierarchical organization of the variables was constructed manually in a flexible manner that allowed for different breaks for different types of bins. Thus for shelter costs at the second level of the hierarchy (number of rooms) homeowner's with or without a mortgage were classified into groups of less than 6, 6-7, and 8 or more while renters were in groups of less than 4, 4, and five or more.

9.3 Stochastic Match

The stochastic match of records was performed at the household level and required only the duplication of FAMEX records. In order to make the fullest possible use of the FAMEX data without having to duplicate SCF records matching bins were created in such a way as to ensure that the FAMEX bin sample size was always smaller than its SCF counterpart. Because the unduplicated host dataset was approximately four times as large it was infrequent that a bin would have to be redefined because the SCF bin had been exhausted before its FAMEX counterpart. The match took the form of a weighted duplication of FAMEX records and forced the FAMEX sample counts within bins to match the corresponding host bin.

In order to perform the duplication a weighted probability of occurrence of household i in bin j is calculated. By multiplying this probability by the desired target bin sample size minus the donor bin sample size an estimate of the number of times a given household should appear in the host dataset is obtained. If the resulting figure is simply rounded or truncated to its integer equivalent, rounding error can produce an incorrect total target bin count. To correct for this error a cumulative total of the target cell frequencies D is calculated (D).

$$D_{ij} = \sum_{k=1}^{i} \left(\left(\frac{W_{ij}}{\sum_{i=1}^{n} W_{ij}} \right) x \left(N_j^t - N_j^d \right) \right)$$

Where: i = the ith household

j = the jth matching bin

W = the weight of the FAMEX donor record

N^t = the sample size of the SPSD target bin

N^d = the sample size of the FAMEX donor bin

Each FAMEX record is then duplicated by the rounded value of the cumulative total minus the rounded value of the previous record's cumulative total plus one. In this way the rounding error is distributed throughout the cell, every FAMEX record is ensured at least one match, and the correct target cell totals are reached.

This procedure serves largely to preserve the weighted distributions of the FAMEX data, at least until SPSD weights are associated with it. The difference between the SCF and FAMEX weights can however create distortions in the matched distributions.

9.4 Evaluation

Several tests to assess the quality of results and assist in subsequent analysis were performed. The distributions of the aggregate expenditures are extremely similar before and after matching. The only real sources of distributional and aggregate difference are attributable to the different (SPSD) weights now associated with the FAMEX data and the minor impact on FAMEX weights of imposing a minimum duplication of one. Benchmark control totals for most expenditure data are not readily available. As such the central test for these aggregate totals were how the post match totals compared to the FAMEX totals. The differences between the individual item totals imputed during the shelter and child care matches were all within five percent. Table 5 presents the results of the expenditure vector match.

Table 5 shows the relationship between the aggregate totals for FAMEX, SPSD, and SPSM modeled variables. The second column shows percentage differences between the pre-and post-matching value of the FAMEX items. As can be seen all of the totals for variables are within a few percent, the differences being largely attributable to the SPSD weights associated with the FAMEX expenditures. Account Balancing Differences are 17.2 percent smaller due to the fact that they are not an actual expenditure but the discrepancy between a family's receipts and disbursements. The third column shows the percentage difference between the FAMEX data and the SPSM modeled and/or imputed variables. The larger differences are due to corrections for underreporting that have been made through the imputation of Green Book distributions.

The degree to which a FAMEX record was duplicated averaged 6 times for all three matches. The maximums duplications were 28, 42, and 51 for the shelter, child care and expenditure vector matches. In 75 percent of expenditure vector matches the duplication was less than 12.

The correlation of host and donor incomes was high with values of .91 and .96 for shelter and child care imputations. The correlation is inversely proportional to the number of bins because of the final sort on income. For this reason the expenditure vector match achieved a weaker correlation (.86). The differences in incomes tended to be the greatest at the tails of the distribution where the most change had been caused in

the host distributions by IPA and high income adjustment. Overall, in 90 percent of cases individual household income differences were within fifteen percent (see Table 5). This fit is especially important to understand because of its effect on various commodity tax model options as well as the apriori relationship between income and expenditures.

Table 5

Expenditure Vector Comparisons, Selected Items

Income/Expenditure Category	FAMEX $ Millions	SPSD/ FAMEX	SPSM/ FAMEX
Food & Non-Alcoholic Beverages	30,805	3.10	
Alcoholic Beverages	4,959	0.38	
Tobacco & Related Products	3,453	3.69	
Men's and Boy's Clothing	4,462	-0.21	
Gross Imputed Rents	19,021	-5.36	
Gross Paid Rent	12,773	4.17	
Electricity	4,226	1.26	
Other Fuels	2,115	8.89	
Durable Household Appliances	3,292	-0.55	
Semi Durables	3,627	-0.98	
Non Durables	4,301	1.22	
Domestic Services	1,121	-9.25	
Other Household Services	2,000	0.40	
Medical Care	1,381	1.65	
Hospital Care	86	-3.46	
Drugs & Sundries	1,657	0.87	
New & Used Automobiles	10,014	-1.53	
Auto Parts & Repairs	4,458	3.63	
Purchased Transportation	3,086	1.43	
Communications	3,583	1.98	
Recreation, Sports, & Camp Equip.	7,514	-4.28	
Books, Magazines, & Stationary	2,261	1.50	
Recreational Services	4,412	0.66	
Jewelry, Watches, & Repair	1,033	-3.41	
Personal Care	2,333	-0.49	
Union & Professional Dues	985	2.75	1.80
Personal Taxes	45,148	-5.19	14.77
Unemployment Insurance Premiums	2,924	0.81	17.25
Retirement Pension Payments	6,108	0.44	18.88
Unallocated FAMEX Items	2,525	8.99	
Net Change In Assets/Liabilities	16,021	-5.49	
RRSP Contributions - Total	3,492	-7.56	36.76
Other Money Receipts	5,612	4.14	
Account Balancing Difference	1,245	-17.20	
Spending Unit Total Income	272,714	-0.87	6.99

Table 6

FAMEX Expenditure Vector Match Income Comparisons

	SPSD/FAMEX Income Percentile Cut-Points						
Quintile	1	5	25	50	75	95	99
1	0.010	0.5985	0.918	0.991	1.055	1.319	1.664
2	0.845	0.887	0.954	0.989	1.021	1.078	1.129
3	0.898	0.938	0.980	1.003	1.030	1.074	1.094
4	0.916	0.947	0.978	1.000	1.022	1.072	1.101
5	0.855	0.900	0.961	0.998	1.037	1.130	1.207
>$80,000	1.003	1.014	1.074	1.181	1.418	2.130	3.418
All	0.555	0.866	0.965	0.999	1.035	1.154	1.572

This table shows the distribution of ratio differences between SPSD income and FAMEX income after the expenditure vector match by income quintiles. Records having over $80,000 income were subset from the fifth quintile. In all quintiles the median ratio difference between pre and post matching incomes was within one percent except for the over $80,000 group. This is because the maximum income on FAMEX is on the order of $250,000 while the maximum on the SPSD is about 11 million due to the high income adjustment. Certain commodity tax model options attribute indirect taxes based on dollars of imputed expenditure and as such the relationship with income should be close.

10. CONCLUSIONS

The Social Policy Simulation Database and Model as described are prototypes of a complex capability for public policy analysis. The process of database creation has undergone two iterations for two different years with a correspondingly massive amount of data. The modeling capability has also been completely developed in at least two hardware/software environments as well as for several major policy options not yet in law. All this has happened effectively since autumn 1984.

In order to test the viability of the SPSD/M idea, a proof of concept, it was necessary to forge ahead making many assumptions and taking short-cuts which made the creation of a testable product possible. This process has had some valuable spin-offs in the form of suggestions for the providers of the source data sets and isolating gaps in the data available for public policy formulation. Furthermore the model has produced results that have already been useful in several instances of policy planning in Canada.

In the short term many methodological refinements of the database creation process will be sought and tested in order to adjust for gaps and inaccuracies in the data. Ultimately further refinements will only be possible due to the collection of more detailed, accurate and timely data.

11. APPENDIX A

Contents of The SPSD

a) Household Structure

Household Weight
Household Sequence Number
Economic Family Sequence Number
Census Family Sequence Number
Individual Sequence Number
Relationship to Household Head
Relationship to Economic Family Head
Relationship to Census Family Head
Tax Filer Status (Imputed)
Common-law Status (Imputed)
Number of SCF Record Clones
Number of UI Record Clones
Number of FAMEX Record Clones
Number of FAMEX Child Care Clones

b) Individual Social Characteristics

Province
Urbanization
Age (Spring 1985)
Sex
Marital Status
Occupation
Industry of Occupation
Years Since Immigration
Labour Force Status (Last Week)
Level of Education
School Type
Educational Status
Weeks Worked Last Year
Weeks Unemployed Last Year
Main Non-Labour Force Activity Last Year

c) Income Components - Market

Employment Earnings
Self-employment, Non-farm Earnings
Self-employment, Farm Earnings
Income from Roomers
Retirement Pension Income
Other Money Income - Taxable
Other Money Income - Non-taxable
Interest Income
Dividend Income
Capital Gains/Losses
Other Investment Income

d) Income Components - Transfers

Family Allowances
Old Age Security Benefits (OAS)
Guaranteed Income Supplement Benefits(GIS)
Spouse's Allowance Benefits
Provincial GIS Top-ups
Canada/Quebec Pension Plan
Other Transfers - Taxable
Other Transfers - Non-taxable
Social Assistance Income
Unemployment Insurance Benefits

e) Income Components - Deductions

Other Employment Expenses
RPP Contributions
RRSP Contributions
Professional and Union Dues
Tuition Fees
Child Care Expense Deduction
Child Care Expenses - Household Total
Business Investment Losses
Carrying Charges
Other Deductions from Total Income
Other Personal Exemptions
Medical Expense Deduction Allowed
Charitable Donation Deduction
Disability Deduction
Education Deduction
Non-capital Losses
Capital Losses
Other Deductions from Net Income

f) Income Components - Tax Credits

Child Tax Credit
Federal Political Contribution
Investment
All Other Federal
All Provincial

g) Income Components - Taxes

UI Contributions
C/QPP Contributions
Net Federal Income Taxes
Net Provincial Income Taxes

h) Household Expenditures

Food and Non-Alcoholic Beverages
Alcoholic Beverages
Tobacco Products and Smokers Supplies
Men's Clothing
Boys Clothing
Women's Clothing
Girls Clothing
Infants Clothing
Footwear and Shoe Repair
Gross Imputed Rents
Gross Paid Rent
Other Lodging
Electricity
Piped Gas
Other Fuels
Furniture, Carpets and Floor Covering
Durable Household Appliances
Semi Durables
Non Durables
Laundry and Dry Cleaning
Domestic Services
Other Household Services
Medical Care
Hospital Care
Other Medical Care
Drugs and Sundries
New and Used Automobiles
Auto Parts and Repairs
Gasoline, Oil, and Grease
Other Automobile Related Services
Local and Commuter Purchased Transportation
Inter-city Transportation
Telephone Communication
All Other Communications
Recreation, Sports, and Camp Equipment
Books, Magazines, and Stationary
Recreational Services
Educational and Cultural Services
Jewelry, Watches, and Repair
Toilet Articles, Cosmetics, Etc.
Personal Care
Expenditure In Restaurants and Hotels
Interest On Personal Loans
All Other Personal Business
Gifts and Donations To Charitable Institutions
Money and Other Gifts To Persons In Canada
Union and Professional Dues
Other Operating Expenses (Non-Profit Institutions)
Personal Taxes
Unemployment Insurance Premiums
Retirement Pension Payments (RPP,C/QPP)
Unallocated FAMEX Items
Net Change In Assets/Liabilities (excluding RRSP)
RRSP Contributions - Total
Other Money Receipts
Account Balancing Difference

i) Memo Items (Income)

Total Employment Income
Total Investment Income
Total Other Market Income
Total Transfer Income
Total Money Income
Total Taxes
Total Disposable Income

j) Housing Characteristics

Tenure (including institutionalized)
Number of Rooms
Number of Bedrooms
Rent Paid
Mortgage Interest
Property Taxes Paid
Dwelling Insurance
Utilities
Repairs and Maintenance
Other Shelter Costs
Market Value of Home
Mortgage Balance Oustanding

k) Claim Data (UI Claimants Only)

Clain Sequence Number (1st or 2nd current year)
Repeater Flag
Initial Benefit Type
Type Change Flag
Weeks of Benefits (current claim)
Weeks of Benefits (in previous 52 weeks)
Weeks of Work (prior to current claim)
Average Weekly Earnings (prior to claim)
Penalty for Voluntary Quit (weeks)
Week Claim Established
Benefits Paid in Calendar Year (1 or 2 claims)
Weeks of Benefits Paid in Calendar Year

SESSION IV: INVITED PAPERS

QUALITY EVALUATION

Chairperson: N.P. Gendreau, Bureau de la Statistique du Québec

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

DATA ON THE ELDERLY — A COMPARISON OF TWO SOURCES

N.J. KOPUSTAS[1]

ABSTRACT

With the growth in the elderly population in Canada, there has been increased interest in timely and detailed data on the elderly. This paper examines two sources of data, the individual income tax file and the Old Age Security file and discusses their strengths and weaknesses as sources of data on the elderly. In addition, methods of combining the data from the two sources to obtain more complete data on the elderly are suggested.

1. INTRODUCTION

Interest in the elderly is increasing and the demand for information on this segment of the population is increasing apace. In an effort to meet this demand, administrative data sources should be considered. The elderly interact with the federal government mainly through two programs: the individual income tax and the Old Age Security programs. These two sources will be examined in this paper in the context of how they can be used to count the elderly, to estimate their income, and to establish their family status (i.e. whether an individual is part of a family or not). In addition, some future directions for possible research will be indicated.

2. THE INDIVIDUAL INCOME TAX SYSTEM

2.1 Population Coverage

In Canada, an individual is required to file a tax return if he or she has a tax liability for the year. A number of deductions increase the threshhold at which an individual becomes taxable. Everyone receives the basic exemption ($4,180 in 1986). The exemption for a spouse with no income was $3,660 in 1986. Individuals 65 years and over are entitled to an age exemption of $2,610. There is also a pension income deduction of up to $1,000 for pension income other than OAS or Canada/Quebec Pension Plan benefits. These deductions total $11,450 for a married individual over 65 with income from a private pension. Excluding the OAS benefit of approximately $3,500 received by nearly all individuals 65 and over, an additional $7,950 from other sources could be received by an individual without generating a tax liability. Other deductions for interest and dividend income, capital gains, medical expenses, and charitable donations could further increase this amount. It should be noted that supplements to the OAS benefits (e.g. Guaranteed Income Supplement) are not taxable.

[1] N.J. Kopustas, Statistics Canada, Small Area & Admin. Data Division, 2306 Main Building, Ottawa, Ontario. K1A 0T6

In spite of the above, approximately 61 percent of individuals 65 and over do file tax returns. This has been relatively constant in the 1980's.

A second reason for filing a tax return is to receive a refundable tax credit. The most familiar is the Child Tax Credit which is payable to mothers. Other credits include a refund of income tax overpayment, a refund of CPP overpayments, and certain provincial tax credits. These generally do not apply to the elderly in a significant way. However, in 1986 a new refundable credit, the Federal Sales Tax Credit was made available to lower income individuals (5 percent of "family income" over $15,000 is subtracted from the credit so that a married couple with income over $17,000 does not receive a credit). Early indications from the 1986 data are that a large number of elderly who were not filers in 1985, did file in 1986 to obtain the Federal Sales Tax Credit. The total number of taxfilers increased by 450,000 or 3 per cent over 1985. If we assume one-third of these are over 65, the coverage of the elderly may now exceed 65 percent.

2.3 Income Information

Most sources of income are taxable, but there are a few that are not and some of these are significant for the elderly. The main source of income for the elderly that is not taxable is the Guaranteed Income Supplement (GIS) paid to lower income elderly. By virtue of this fact however, very few of the GIS recipients file tax returns in any case. Other kinds of income that might be received by the elderly but are not taxable include disability allowances, war disability pensions, veterans' allowances, and the refundable tax credits. In summary, a number of transfer payments are not included in taxable income and therefore the income of the elderly is under-reported. Because some of these sources must be reported in the calculation of the Federal Sales Tax Credit, the coverage of income for the elderly in the tax system should improve.

2.4 Family Composition

Experimental family data has been generated from the Individual Income Tax data by Statistics Canada (Auger, 1987). The concept used is that of the Census or nuclear family consisting of a married couple with or without never married children or a single parent with never married children. The family data are created through a series of steps that identify and link the members of a family that file tax returns including spouses and children under 30. Non-filing family members are imputed (again spouses and children), using the structure of personal exemptions, family allowances received, child tax credit amount, and information on other tax credits.

The results from the 1982 family data for the elderly are compared to the 1981 Census data in Table 1.

Table 1

Population 65 and Over by Family Type, from 1982 Tax and 1981 Census, Canada

Family Type	Tax	Census	Ratio
	(000s)	(000s)	
Non-Family Persons	563	860	0.65
Family Persons	1069	1501	0.71
Husband-Wife	1051	1191	0.88
Single Parent	18	310	0.06
Total Population	1631	2361	0.69

In the family data, an individual is either a taxfiling family member or an imputed non-filing family member. The tax system allows exemptions for several types of dependants including a spouse, children under 18, children over 18, and other dependants under certain conditions. In the current family system, spouses and children are imputed, but not other dependants. Some spouses 65 years of age and over are imputed, raising the coverage rate from 61 percent for filers to 69 percent for filers and dependants. Most elderly dependants would be claimed as other dependants under "additional personal exemptions" on Schedule 6 of the tax return. (See Appendix 1 for a sample of Schedule 6 from the tax return.) However, the current family system does not account for these exemptions. This would be difficult to do for a number of reasons. Unlike the other exemptions where additional information is available upon which to make assumptions about the age and sex of the dependant, no information about the Schedule 6 dependents is available except the total amount claimed for all such dependants. The number of such dependants is not even available. None of the information on date of birth and relationship to taxfiler is captured in the processing of the tax return. Very broad assumptions could be made but the dependants in this category cover a range of dependants that can come from different generations and therefore make determination of age very difficult. The number of taxfilers claiming additional exemptions in 1986 was 3.2 million. How many dependants this represents and the proportion that are 65 and over is impossible to determine.

In a Census family, if the parent files a tax return and the children are either dependants or file their own returns, the current family system can usually construct the appropriate family. On the other hand, if an unmarried child over 30 years old files a tax return and an elderly parent in the same dwelling also files a return, the current family system will construct two sets of non-family persons rather than a single-parent family. If the elderly person does not file, he or she will be lost completely to the system. This would account for the very low coverage of elderly single-parent families.

In summary, the tax data can be used to account for 69 percent of the elderly population, and provide income information for the 61 percent who file tax returns. These coverage rates will increase starting in 1986 with the introduction of the Federal Sales Tax Credit.

Husband-wife families with at least one spouse 65 and over can be identified where one of the spouses files a tax return, but non-filing dependant elderly are not identifiable leading to lower coverage of non-family persons and single-parent families in this age group.

3. THE OLD AGE SECURITY PROGRAM

3.1 Introduction

The Old Age Security program is a virtually universal program for individuals aged 65 and over. A monthly benefit is paid to everyone who conforms to the residency requirements (generally 10 or more consecutive years in Canada) and who applies for the benefit. An additional benefit, the Guaranteed Income Supplement, is paid to lower income elderly and is based on the previous year's income.

3.2 Coverage

The OAS program covers 96 per cent of the population aged 65 and over. People who do not have the residency requirements and people who have not applied for the benefits are not included. It is possible therefore to estimate the population 65 and over quite well from these data.

3.3 Income

Income information for the previous year is required when an individual applies for the Guaranteed Income Supplement. The information required follows the income tax definitions for the most part, except that OAS payments and Family Allowance payments are excluded, and Workers' Compensation benefits are included. To obtain an equivalent total income these adjustments must be made. This is not operationally straight-forward since a given OAS/GIS record contains benefit amounts for the current year, but income information for the previous year. Either data for an individual must be obtained for two years, or assumptions about his status in the previous year must be made (e.g. received the same GIS as the current year).

3.3 Family Status

There is very little demographic information on the OAS data other than age and sex. Marital status is used in determining whether the account is a stand-alone or a joint account. A joint account is created when both spouses are 65 or over and are receiving OAS benefits. If only one spouse is 65 or over the presence or absence of the other spouse is not indicated and therefore no information on that spouse is available. If a person applies for the Guaranteed Income Supplement, the income of the spouse must be included. In this case, the presence of a spouse could be detected. No indication of the existence of dependants or dependancy on others is given. Husband-wife families with both spouses 65 and over can be estimated, but not other types of family relationships as no information on children is available.

4. COMBINING TAX AND OAS INFORMATION

4.1 Introduction

The income tax system does not cover the low income elderly and therefore this source of data is not available for approximately 40 per cent of the elderly. However, this is the very group who qualifies for and applies for the Guaranteed Income Supplement to the Old Age Security program and supplies income information to Health and Welfare Canada to this end. If the two data sources, tax and OAS, could be brought together, it seems likely that fairly complete income information could be obtained. An experimental linkage of the two files was carried out for files containing 1984 income data for residents of British Columbia.

4.2 Coverage

As expected, 96 per cent of the elderly are accounted for by the OAS data and 64 per cent of the elderly filed tax returns in British Columbia in 1984. The proportion linked was 61 per cent, leaving 3 per cent of the elderly filing tax returns but not recieving OAS benefits. Thus the combined data accounts for 99 per cent of the population 65 and over. It should be emphasized that this result holds for one province at one point in time and may not be a general outcome. Other provinces with different age structures and immigration patterns may show different results.

4.3 Income

From the tax data income information is available for 64 per cent of the elderly, of which 15 per cent are receiving GIS, and 49 per cent are not receiving GIS. On the OAS data income information is available for virtually all of the GIS recipients representing 41 per cent of the elderly. Taking the 49 per cent not receiving GIS from the tax data, and

the 41 per cent receiving GIS from the OAS data, income information is available for 90 per cent of the elderly.

Table 2

Counts of Elderly and Income Coverage, British Columbia, 1984

		Income Information						
Status	Number	on %	OAS	on %	Tax	on %	Either	%
No GIS	207,141	58	227	0	174,027	49	174,254	49
OAS and GIS	148,926	42	144,132	40	52,780	15	146,154	41
Total	356,067	100	144,359	41	226,807	64	320,408	90

(All percentages are based on the total of 356,067).

4.4 Family Composition

Since the experimental linkage was based on individual records and not the family data described above, no additional information was gained from this linkage concerning family structure and family type for the elderly.

5. FUTURE WORK

This paper has identified a number of issues that need to be addressed in order to maximize the available information from administrative records on the elderly.

The characteristics of the subset of the elderly not covered by the OAS program should be determined in order to improve any population estimation technique.

Much work needs to be done on the reconciliation of income information between the tax data and the OAS data. Although the sources of income reported are nearly identical, some adjustments must be made. There is also the problem of combining the OAS payments and income reported for the same year within the OAS system. In addition, the impact of non-taxable sources of income for the elderly needs to be studied.

To improve family information, the group of dependant elderly who do not file tax returns needs to be identified. The introduction of the Federal Sales Tax Credit in 1986 will bring part of this group into the tax system. The 1986 data became available in October, and a study is underway to assess the impact of the Credit program on the tax data.

A longer term solution to improving family information on the elderly is to capture the data collected on Schedule 6 of the tax return.

6. CONCLUSIONS

It is clear that counting the elderly is possible using the OAS data with a very small adjustment on the order of 4 per cent. Characteristics of the population are limited to age and sex. Marital status reflects only the presence or absence of a spouse. Geographic detail based on the postal code is possible depending on how the coverage adjustment is incorporated.

Income of individuals can be obtained for 90 per cent of the elderly population through a linkage of the OAS and Tax files. If the GIS recipients could be identified on the tax file and deleted, the linkage would not be necessary. In this case the two files would be complementary. By taking the income information on GIS recipients from the OAS file and the income information on non-GIS recipients from the tax file, the most complete information could be obtained.

Family information is more problematic. Because the OAS file contains no dependancy information and the information on the exemption category in which most dependant elderly are included is not available on the tax file, family types other than husband-wife are difficult to estimate.

Overall, the initial findings in terms of population and income coverage of the elderly from administrative sources are encouraging, and as the tax system moves toward a tax credit system, family data on the elderly from administrative sources should continue to improve.

REFERENCES

Auger, E. (1987). Family Data from the Canadian Personal Income Tax File, Statistics Canada (unpublished).

Podoluk, J. (1987). Administrative Data as Alternative Sources to Census Data, Statistics Canada (unpublished).

Selley, O. (1987). Pilot Study — Microrecord Linkage of the Tax and Old Age Security Records for British Columbia, Statistics Canada (unpublished).

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

A TWO-STAGE SURVEY: THE PERMANENT SAMPLE OF SOCIAL INSURANCE BENEFICIARIES IN FRANCE

ANDRÉE MIZRAHI and ARIÉ MIZRAHI[1]

ABSTRACT

In France, 99% of the population is protected by one of the country's compulsory Health Insurance plans. As a general rule, medical patients in urban centres pay for their care and then obtain reimbursement for the amounts payable by Health Insurance. When patients are hospitalized, the insurers pay the hospital facilities directly for the portion of the expenses covered. A permanent random sample (1/1200) of social insurance beneficiaries has been monitored since 1977, and will involve approximately 40,000 persons in 1989. Routinely collected administrative data on the exact nature and cost of the health services used by each beneficiary will be completed by a survey of the beneficiaries themselves (one-fourth of the sample each year). The survey deals with socio-economic factors; health protection; illnesses and disabilities; and opinions on the health care system. A record of expenditures for medical services used over a three-week period, reimbursed or not, is attached to the survey.

1. HEALTH INSURANCE IN FRANCE

1.1 The Various Plans

In France, Social Security plans protect approximately 99% of the population - in one way or another and in varying degrees - with respect to health care.

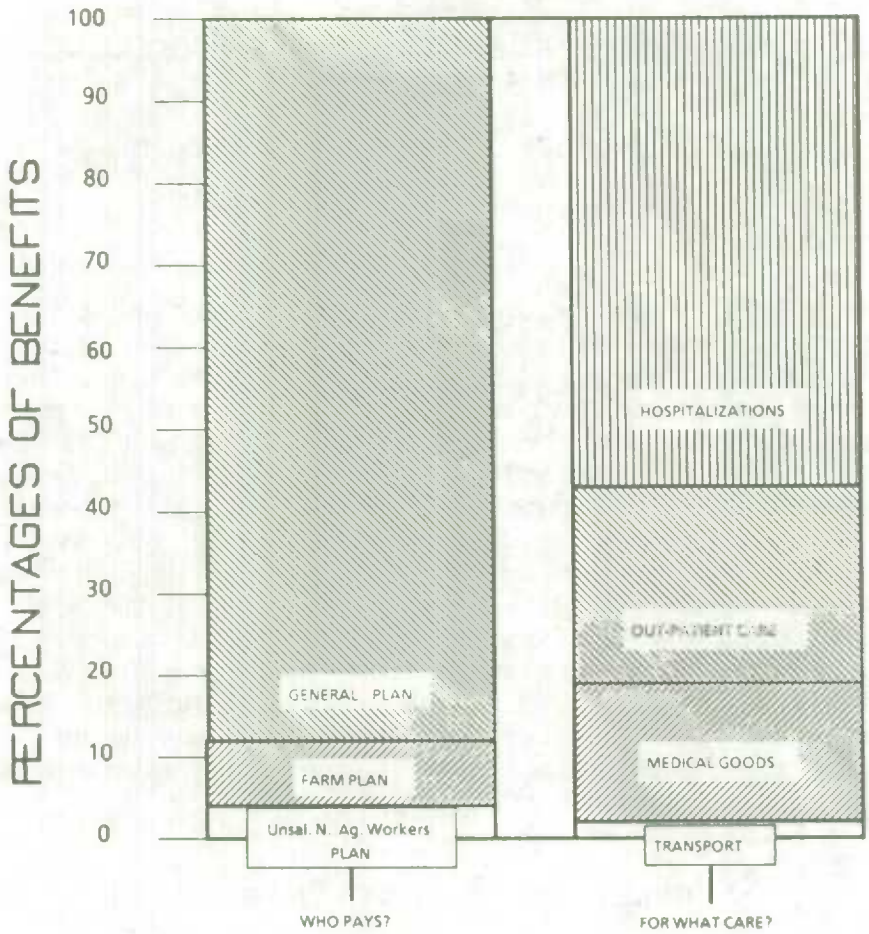
This compulsory coverage, which operates on the basis of occupational activity is provided essentially by health, maternity, and industrial accident insurance under three plans: the Régime Général des Salariés (general plan for wage-earners) (87% of benefits paid); the Régime Agricole (farm plan) (9% of benefits); and the Régime des Travailleurs Non-Salariés des Professions Non-Agricoles (plan for unsalaried workers in non-agricultural occupations) (4% of benefits).

Contributions determined in proportion to salaries (or occupational incomes for unsalaried workers) provide 94% of health insurance funding. In return, health insurance covers the cost of 76% of health services: 86% of hospitalization costs; 66% of out-patient costs; and 61% of the cost of medical goods. (See Graph No. 1).

[1] Andrée Mizrahi and Arié Mizrahi, Centre de Recherche d'étude et de documentation en économie de la santé 1 Paul Cézanne Street, Paris, France, 75008.

Graph No. 1
SOCIAL INSURANCE BENEFITS

FRANCE 1986



Ref: OTT87.DOC - A Two-stage Survey

8 January 1988

1.2 Methods of Assuming Costs

The Health Insurance system operates on the basis of rates set by public authorities (for hospital facilities, pharmaceutical products, lump-sum care for the elderly, and so forth), or negotiated under the form of quadrennial agreements with representatives of occupational groups in the health sector (physicians, dentists, physiotherapists, and so forth).

Generally speaking, the insured person makes direct payments to the physician, dentist, medical auxiliary, laboratory, pharmacist, and so forth, and is subsequently reimbursed by his/her Caisse Primaire d'Assurance Maladie (primary health insurance fund). For the Régime Général (general plan), there are 129 primary funds located throughout the country. Reimbursements cover 75% of the rates for physicians' and dentists' fees, 65% for the amounts charged by medical auxiliaries and laboratories, and 40%, 70% or 90% for pharmaceutical products.

Both active and retired workers are insured: they eventually receive reimbursement for personal health care costs and for those of dependants who are officially described as dependants entitled to benefits under their names.

In the main, Social Security pays hospital facilities directly for their services; beneficiaries pay only the amounts payable by them, 20% of the cost or a lump sum of 25 francs per day of hospitalization.

Health Insurance covers 100% of the charges, however, in certain cases involving certain types of care: those related to thirty particularly long and costly ailments (such as malignant tumours, Parkinson's disease, and hemophilia); maternity, relatively major surgery; hospitalization beyond thirty days; industrial accidents; and so forth. Coverage for 100% of health care costs is also provided for certain persons, including: beneficiaries of the Fonds National de Solidarité (national solidarity fund); veterans; victims of accidents in the workplace; and certain disabled persons. Ten percent of those protected (the insured and those entitled to benefits under their names) reportedly benefit from at least one of the above-mentioned measures.

In reality, the situation is more complex because the rates are not always the same as the amounts actually charged. Thus, health costs for families are higher than the amounts payable by them according to the regulations, in so far as nearly 30% of physicians, most dentists when supplying dental protheses, certain medical auxiliaries, opticians, and other care providers officially charge more than the set rates.

Lastly, in France in 1986, 76.7% of medical expenses were paid by Social Security and 14.6% were assumed by the population. Complementary insurance plans covered 7.2% of costs (4.3% covered by mutual insurance companies and 2.9% by personal insurance), reimbursing a portion of the expenditures not payable by Social Security. Medical Aid (an assistance program for persons with insufficient resources) assumed 1.5% of such costs.

1.3 Social Security Records and Statistics

The Health Insurance system continually generates, as a by-product of its administration, a great many statistics that constitute records of all its operations and are widely used by administrators in the medical sector; management and labour; economists; and so forth. These data form, in particular, the basis for the development of the Comptes Nationaux de la Santé and the Comptes Nationaux de la Protection Sociale (national health and social welfare accounts), which are used extensively for the development of health policies on both the national and international levels.

The primary function of the records held by the Health Insurance system, however, is to make possible the proper payment of benefits both to the insured, when direct payment is made for health care, and to the providers (or distributors) of care, in cases involving direct payment by insurers.

Usable data thus relate to the insured (registration record: last name; first name; address; age; sex; and nature of current entitlements); to health care providers (providers' records: identification; specialty; and right to exceed rates); and to amounts payable (benefits records: identity of the insured and the beneficiary; key letter and code assigned to care, according to the schedule of professional services; dates; and so forth).

These records are kept by health insurance data processing centres located throughout the country.

Data required for payments to beneficiaries and care providers, such as names and addresses of the insured and amounts paid, are recorded in great detail; this is much less true with respect to data not directly required for payment. Thus, hospitalization, which accounts for more than half of all medical expenditures, is not always recorded in full detail in the files of the persons concerned. Since hospital facilities are financed by a general budget, these data are not absolutely necessary for payment of benefits. This is also the case for lump sums which are paid to institutions that provide residential care for dependent elderly persons, and that are reimbursed by health insurance.

Updating records occasionally requires rather lengthy periods of time, especially for data not affecting the payment of benefits and for cases that involve deceased beneficiaries, persons reimbursed at former addresses, and names which appear in several files. (In any event, the system makes repeated reimbursement for the same service impossible.) As a result of these extended time periods, a certain amount of doubt exists as to the number of persons protected by the various plans.

1.4 The Permanent Sample of Social Insurance Beneficiaries

Furthermore, since health insurance data is highly complex and recorded in many different places, it is difficult to establish statistics with respect to individual beneficiaries. With a view to offsetting these deficiencies, the statistics department of the Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS) (national health insurance fund for salaried workers), in co-operation with the Centre de Recherche, d'Etudes et de Documentation en Economie de la Santé (CREDES, formerly CREDOC) (centre for research, studies and documentation in health economics), established ongoing monitoring of a representative random (1/1200) sample of social insurance beneficiaries.

This sample group of social insurance beneficiaries constitutes one of the major tools available to the Social Security system for analysing health insurance expenditures from the point of view of the users of medical services (micro-economic) and for measuring the effect of changes in laws and regulations governing health insurance. The sample group established in 1977, through the voluntary efforts of seven of the 129 primary health insurance funds, was gradually expanded when its effectiveness was demonstrated. It will involve approximately 32,000 persons, or 82% of the selected field, in early 1988, and is expected to reach its full size in 1989, when the data available will involve approximately 39,000 persons protected by the Régime Général de Sécurité Sociale (general social security plan), or a 1/1200 sample.

The sampling method chosen (all those whose Numéro National d'Identité (national identification number) confirms a certain degree of suitability) ensures the automatic renewal of the sample.

All data concerning the insured persons thus selected and those entitled to benefits under their names (identification and benefits) are routinely recorded (by data processing systems, as they are generated) and centralized to serve in the development of statistical applications and economic analyses.

The features of the health insurance records are found in those of the sample group of beneficiaries, and data which are highly significant for socio-economic analysis of consumers' behaviour patterns (occupation, supplementary coverage, level of education, and so forth) are not contained in these records. It was therefore decided to add information, to be obtained from the beneficiaries themselves, to the data already available: this stage of the survey, which involves contacting households directly, is called the **Survey on Health and Social Security.** All data obtained from both stages are indicated in Table 1.

Table 1

Permanent Sample of Social Insurance Beneficiaries
Source and Nature of Data

Source:

SAMPLE OF SOCIAL SECURITY FILES	DIRECT HOUSEHOLD SURVEY
Field:	**Field:**
Insured persons + those entitled to benefits under their names	Insured persons + those entitled to benefits under their names living under the same roof + other household members + persons not living under the same roof entitled to benefits under the insured person's name
Nature of data:	**Nature of data:**
	Demography
Age and sex	Age and sex (including persons not living under the same roof entitled to benefits under the insured person's name)
	Social Security
Insured person's Social Security plan in full	Health Insurance plan and supplementary protection for each household member
	Socio-economics
	Activity/occupation and level of education for all household members and persons not living under the same roof entitled to benefits under insured person's name
	Epidemiology
	Disease rate; disability
	Medical services
Type of provider, location, nature of care (specific type of medical services) and level (key letters), date, rates, amounts spent, benefits paid **(data routinely recorded)**	Type of provider, location, date,nature of care, expenditures for medical services used **(information obtained during three weeks only)**
	Opinions
	Future of Social Security system, measures proposed

2. STAGE TWO: HOUSEHOLD SURVEY OF SAMPLE GROUP OF SOCIAL INSURANCE BENEFICIARIES

Stage Two, the household survey, will consist of a survey conducted each year involving a quarter of the permanent sample of social insurance beneficiaries, selected randomly, so as to cover the entire sample in four years.

2.1 Pilot survey

The survey method was developed on the basis of the results of a pilot survey conducted in May and June 1987, during which five methods were used in an experimental manner. These methods differ with respect to type of contact (telephone calls or visits by interviewers); length (normal or abridged form); number of contacts (two or four telephone calls); and composition of the network of interviewers (employees of a professional survey organization, Social Security administrative officers, or CREDES researchers).

Three of the five methods differ only in the type of interviewers involved:

a) - professional interviewers;
b) - CREDES researchers;
c) - Social Security officers;

} In these three cases, the survey is described as "normal" and consists of a minimum of four telephone calls.

d) - a telephone survey conducted by professional interviewers, described as "abridged" and consisting of a minimum of two telephone calls; and

e) - a household survey described as "face to face" conducted during two visits by professional interviewers.

When one and the same (normal) survey is used, the refusal rates noted by non-interviewers (CREDES researchers and Social Security administrative officers) are significantly higher than those noted by specialized personnel.

Return rates for health care records and "health questionnaires" are slightly lower when telephone calls are used than when visits are made. On the other hand, positive response rates for the initial telephone call are significantly higher than those for the initial visit.

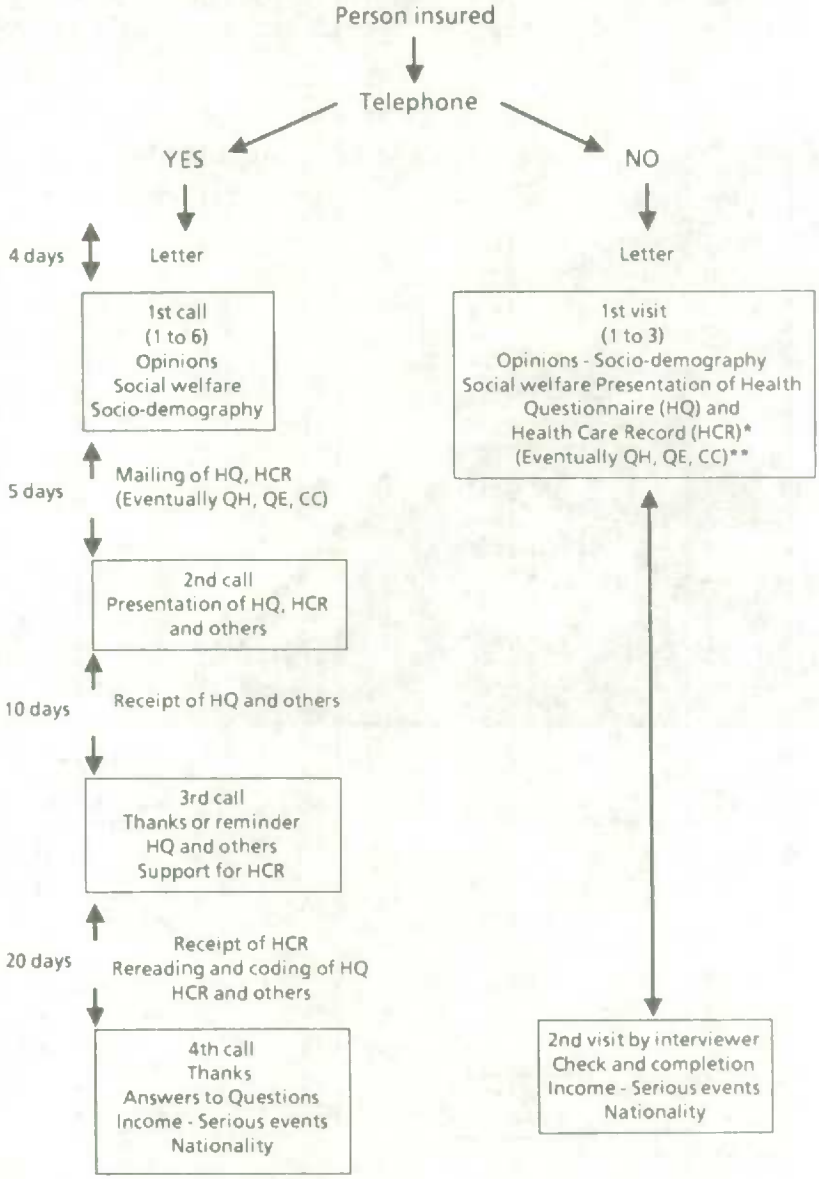
2.2 The Method Selected

Consequently, organizations specialized in conducting household interviews will be given responsibility for field operations.

Contacts will be made by telephone and mail for all households with telephones. Other households will be visited by interviewers. The telephone questionnaire will be divided into two (not three) parts, to be administered during the first and last calls, while two support calls are planned during the period in which the health care record is kept.

For households without telephones, two visits are planned: the first to administer main and secondary questionnaires and deliver health care records; the second to collect health care records and check the consistency of all information provided. Graph No. 2 shows the sequence in which these operations take place.

Graph No. 2
PERMANENT SAMPLE OF SOCIAL INSURANCE BENEFICIARIES
SEQUENCE OF EVENTS DURING HOUSEHOLD SURVEY STAGE



Ref: OTT87.DOC - A Two-stage Survey

8 January 1988

* QS = Questionnaire Santé; CS = Carnet de Soins. - Translator's note.
** Abbreviations QH, QE and CC impossible to elucidate. QH and QE may refer to 'other questionnaires' mentioned in text; CC may refer to another type of 'carnet' or record. - Translator's note.

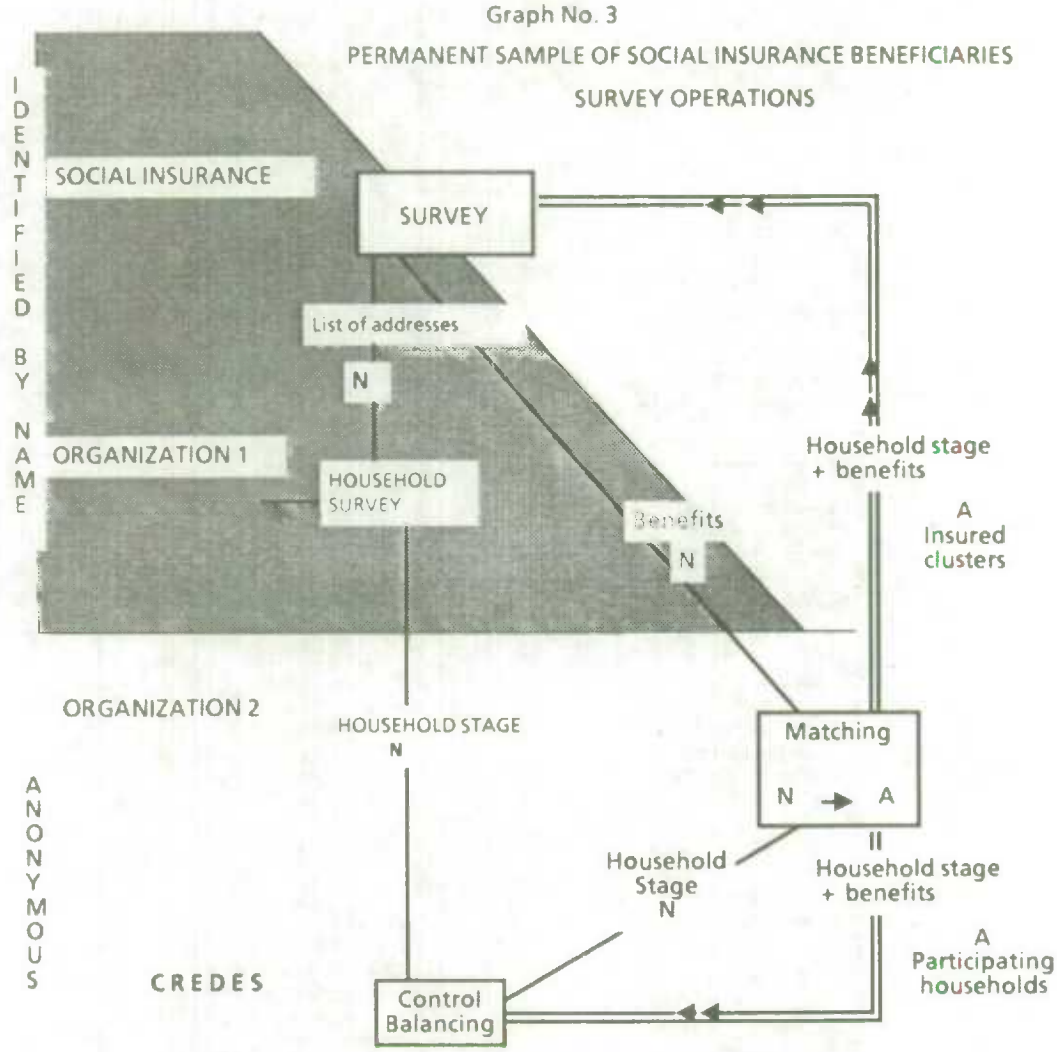
All documents are reread and coded as they are completed or received, with a view to identifying as many inconsistencies and inaccuracies as possible; interviewees are questioned during the next telephone contact regarding the points thus identified. This procedure considerably improves the quality of the data obtained.

In order to comply with restrictions protecting privacy and public freedoms, all these data are handled in a strictly anonymous manner. This condition causes no difficulty in either of the two stages: the sample of Social Security records or the direct household survey. Double-blind matching of these two sources is done by a third organization which is provided with the records after they have been rendered strictly anonymous. (See Graph No. 3.)

Graph No. 3

PERMANENT SAMPLE OF SOCIAL INSURANCE BENEFICIARIES

SURVEY OPERATIONS



Réf. OTT87 DOC A Two-stage Survey

8 January 1988

3. CONCLUSION

Many problems must yet be solved: statistical processing of longitudinal micro-economic data (benefits); consolidation of observations relating to successive periods (breakdown of the sample into four sub-samples, each observed once during the period); proper combination of data from different sources; and corrections of eventual inconsistencies.

Nonetheless, because of the wealth of information obtained, this work opens the way for highly significant research. We hope that the survey will enable researchers to monitor the development of health-related events over time, and, in particular, to follow the development of the influence exercised by socio-economic standards and disease rates

on the level and structure of the use of medical services, thus making micro-economic series available for use in the medical sector.

REFERENCES

C.N.A.M.T.S. (1986). Statistiques des Régimes d'Assurance Maladie en 1985. Département Statistique.

C.N.A.M.T.S. (1982). Qui consomme quoi ? Département statistique, Paris. Mise à jour des résultats issus de l'échantillon permanent d'assurés sociaux, Décembre 1984.

S.E.S.I. (1987). Comptes Nationaux de la Santé 1984-1985-1986. Ministère des Affaires Sociales.

Mizrahi, Andrée and Mizrahi Arié. (1978). Méthode de sondage: Enquête permanente dans les dossiers de sécurité sociale. Credoc, Paris.

Volatier, J.L..Enquête sur la santé et la protection sociale. Methodologie de l'enquête pilote. CREDES. (to be published)

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

SYMPOSIUM ON STATISTICAL USES OF ADMINISTRATIVE DATA
CORPORATION INCOME TAX RECORDS USED FOR TAX POLICY ANALYSIS

F. HOSTETTER, C.D. McCANN and B. ZIRGER[1]

ABSTRACT

Over the past several years, tax and financial data items have been extracted from corporation income tax returns on a sample basis and used by tax policy analysts in the Department of Finance to study the effects of hypothetical changes in tax policy on tax revenue and on tax status. While income tax records are an appropriate source of this data, quality problems can arise when records created for administrative purposes are used for statistical purposes. Difficulties can arise from the filing and assessing processes, non-standardization of both tax reporting and financial reporting by corporations, data definitions and interpretation and the data extraction process. Quality assurance activities have been implemented with success, to identify and correct errors associated with these difficulties.

1. INTRODUCTION

In this paper we describe statistical data collection operations in National Revenue Taxation (NRT) using corporation income tax records. As well, we outline present and planned data quality assurance activities. The first part of this paper provides some background information on the client's priorities and how these are reflected in the sample. We then describe the corporate environment in which the data collection operations are situated, the relevance of this environment to the statistical use of corporation income tax records, and data collection operations and the associated quality assurance activities. The paper finishes with some possible future directions for quality assurance of sampled corporation tax data. We start with a description of the 'top-down' view of the end uses of the data collected.

2. BACKGROUND

Ensuring the quality of corporation statistical data is a small part of a much larger, ongoing process culminating in Parliament making changes to the Income Tax Act. Such changes to the Act deal specifically with taxation of corporations and result from the ongoing evaluation of both existing and proposed tax policy measures. This evaluation is carried out in the Department of Finance and is based on estimates of the revenue effects of such tax measures, among other considerations. The principal technique used to estimate these revenue effects is micro simulation using NRT's Corporation Tax Model.

[1] F. Hostetter, C. McCann and B. Zirger, Revenue Canada, Taxation, Room 303 MacDonald Building, 123 Slater Street, Ottawa, Ontario, Canada K1A 0T6.

The micro-units on which simulations are based are sampled corporation tax records, which are drawn from a sampling frame defined in terms of NRT's administrative file for corporations. Data obtained from this administrative are supplemented with additional data extracted from the sampled corporations' financial statements and tax schedules. These data are then integrated into a cross-sectional representation of corporate taxfilers for a particular tax year. This representation is a computer-readable file called the Corporation Sample File and becomes the primary data input to the tax model.

National Revenue Taxation started these statistical data collection operations in 1979. They were based on a specification by our client, the Department of Finance, of their priorities for corporation tax modelling and the data required to meet those priorities.

The client's priorities addressed the requirements for tax policy analysis with particular regard to corporate ownership, tax status and industry. Corporate ownership effects eligibility for the numerous tax provisions for small business. Tax status is relevant in estimating the revenue effects of a tax change for taxable corporations. Firms within a particular industry are typically eligible for particular tax provisions. As well, Department of Finance analysts wanted a classification by geographical jurisdiction to ensure a suitable quality of estimates of tax changes by regions. As the size of the corporation was frequently related to the number of items of interest for tax policy simulations, it was felt that corporation size should also be a priority. In summary then, the client's priorities were:

Regional Comparisons
Specific Industrial Sectors
Large and Small Businesses
Tax Status and Ownership
Estimation of Simulation Effects on Tax

Based on these specifications, a sample was drawn from a stratified corporate population using simple random sampling. These strata were defined by 13 geographical or tax jurisdiction, 25 industrial groups, 6 variable asset size ranges, 2 tax status categories. The present corporation sample consists of about 16,000 corporations for which there are a possible 1,000 data items each. More than half of these data items are obtained in the statistical data collection operations, with the balance obtained from NRT's corporation administrative file. By way of example, the number of possible data items per corporation obtained in these statistical data collection operations for the most recent year is as follows:

Balance Sheet		56
Income Statement		44
Comparative and Miscellaneous Finance Data		21
Subcodes		8
Sources and Uses of Funds		22
T2S(1)	Reconciliation Statement	47
T2S(3)	Dividends Received	4
T2S(4)	Continuity of Losses Carried Forward	20
T2S(6)	Disposition of Capital Property	8
T2S(7)	Investment Income	8
T2S(8)	Capital Cost Allowance	164
T2038	Investment Tax Credit	59
T2S(12)	Continuity of Earned Depletion Base	36
T2S(13)	Continuity of Reserves	2
T2S(15)	Pension Contributions	4
T2S(27)	Manufacturing and Processing Profits	11

The manual process to extract these data items is continuous throughout the year. It is normally carried out by permanent clerical personnel from the staff of the Ottawa Taxation Centre. The Statistical Services Division in Head Office provides functional direction. This functional direction includes:

- resourcing the operation,
- specifying production targets and productivity rates,
- providing the work instructions,
- providing the data transcripts,
- streaming the workload and ensuring it is available when required,
- monitoring production and resource utilization, and
- monitoring data quality

Prior to describing how data quality is controlled, we will digress briefly to discuss some key considerations in using administrative data.

3. USING CORPORATE TAX RECORDS FOR STATISTICAL PURPOSES

The Canadian tax system is based on voluntary compliance whereby taxfilers determine their own tax liability. Within that system priority tasks for NRT include ensuring the accuracy of taxfilers' assessments, and treating taxfilers fairly and professionally.

In an organization whose primary business is tax administration, data collection for statistical purposes is a secondary consideration. The vast majority of NRT employees administer the tax system. As such, they are typically engaged in assessing, auditing or collecting tax. Many of them have an accounting background. In contrast, the handful of NRT statisticians are engaged in analysing and reporting on such matters as the numbers of taxfilers having particular individual or corporate characteristics, and the tax these filers pay under particular financial circumstances.

In other words, the corporate environment, or culture, into which statistical data collection operations are introduced is dominated by the requirements and priorities of tax assessing and collecting. The environment, and the tax system itself, are important factors influencing both the quality of statistical data and the actions that may be needed and can be taken to improve that quality. The generally positive impact of these matters in NRT becomes evident when considering survey coverage, data access, survey response and control.

In terms of survey coverage, the NRT corporation administrative file can be viewed as a complete census of the corporate population. It contains a magnetic record of each tax year's data on each corporate filer of the T2 form in Canada. All incorporated businesses, regardless of whether they are taxable or active, must file a T2 return. Thus there are about 700,000 corporations represented on this file for a single tax year.

The data items captured and stored for each filer on this magnetic record result from the administrative procedures in the Department, and the data items are limited accordingly. Most of the detailed data items available from the tax schedules and financial statements normally attached to the T2 return are not captured on this administrative file. Because the corporation administrative file is very large but limited in its details, it can not be used to good advantage to perform micro simulations. Consequently, a sample is drawn to address this need.

Access to the administrative file and to the T2 returns by NRT statisticians and other personnel for statistical purposes is fairly straightforward. Of course, their access is secondary to the requirements of NRT's assessors and auditors. Access by other persons,

including statisticians in other departments, is strictly limited by the provisions of the Income Tax Act designed to preserve confidentiality and privacy of taxfiler data.

The issue of survey response is likewise not a significant problem. Since the sample is drawn from a target population which is a subset of the administrative file, virtually all reporting units selected are obtained. A small problem occurs infrequently when the T2 return being sought is being used in an audit. However, this is largely eliminated by photocopying the original T2 return, selected for inclusion in the sample, once it has been computer-assessed. If a reporting unit ignored the law and did not file a return for a particular year, then it would not be in the administrative file. Consequently, it would not be available for selection in the sample survey. This potential problem is also minor since filing compliance is high.

Lack of control over reporting units, data items, response and access is likewise not problematic. It should be underlined, though, that the administrative system for assessing and processing corporate tax returns always has priority over statistical uses of such data.

In summary, coverage, data access, survey responses and control are relatively problem-free, largely because both the administrative and the statistical users of corporate tax records belong to the same department and share a common corporate culture.

The operating requirements of NRT and its corporate culture shape the activities carried out to forestall, identify and correct errors in corporate tax records used for statistical purposes. The following section describes those activities.

4. QUALITY ASSURANCE ACTIVITIES

We initially wanted to describe a possible future statistical quality control project applied to corporation tax and financial data. That initial desire led us to question the definition of quality and how it might be possible to measure the quality of corporation statistical data. However, since there are about 700 items of data involved, it would seem impractical to focus on a measurement of the quality of each one of those data items. We therefore settled on a definition of quality which recognizes that quality is the sum of all characteristics that determine the acceptance of a product in relation to the intent of design, specifications and user priorities. The definition seems to be clear enough, but it does not tell how to measure quality. Rather than describing how we might like ideally to measure quality, we opted for describing the several existing quality assurance activities related to corporation statistical data collection.

The intent of these quality assurance activities is to ensure that the product meets the specifications established on the basis of client requirements. The specifications for the data to be collected are detailed in NRT's Corporation Data Analysis Manual. This manual sets out the procedures used by the data analysts in the Ottawa Taxation Centre to extract data from the corporation tax returns and the financial statements attached to them.

There are many sources of potential error that can distort this accuracy. The first source of error could be the taxfiler himself. The taxfiler may make an error in filling out the form or in filling out the financial statements. When the T2 tax return is sent to NRT, it is assessed to verify the accuracy of the data related to tax fields. The verification of the data by assessors is done within assessing tolerances, and hence the data precision needed for statistical purposes may be absent. Although the assessors collect some statistical data on behalf of a number of different users in NRT and in Statistics Canada, these statistical data are not subjected to any automatic editing, and also may not be free of error. The Department's main administrative file of corporation data is built up from the transactions prepared by the assessors at this point.

Inconsistent interpretation of the specifications or workload pressures may lead to errors in the data extracted. An emphasis on quantity of production can also sometimes adversely affect the quality of the data extracted.

Once the data have been analysed by the corporation data analysts, reviewed and transcribed onto the transcript, the transcript is then sent for key editing from the transcript into a computer. Errors may be made at that point as well.

Later on in the process we find that some errors are introduced as a result of the implementation of the sample design. Sometimes the corporation returns selected for inclusion in the sample are replaced in the wrong strata. As a result, the observations on the file are inappropriate or inconsistent with the intent of the sample design.

We carry out seven activities to compensate for these sources of potential error and to ensure the quality of corporation statistical data. In the Ottawa Taxation Centre the activities consist of a review of the administrative file data by the data analysts, a transcription review, data capture checks, and automatic edit checks and manual correction. In the Head Office Statistical Service Division we assess the quality of the data after it has been collected, identify any problem areas in the corporation data analysis process, and decide on corrective measures. The final two quality assurance activities consist of applying edit checks, which are added to the system throughout the year, to all observations in the sample and reconciling summary statistics against other sources of data. Each of these seven activities helps to improve the quality of the data which are released to the Department of Finance and used in tax policy analysis. Each of these activities are described below.

Review of the administrative file data is performed by the corporation data analysts at the Ottawa Taxation Centre. For each tax return selected in the sample, about eighty data items are taken from the Department's main administrative file and printed on a special transcript. These data items are cross-checked against the source document - the T2 return and financial statements. If there is a difference, the corporation data analysts reconcile the difference. If the administrative data field is in error, a correction is made to it which then goes on to the corporation sample file before it is released.

Once the transcription is completed, an experienced, senior data analyst does a manual transcription review. Errors identified at this point are returned to the analyst and corrected before the transcription is keyed into the computer. The percent of review of a particular analyst's work varies according to the past performance of the analyst, the analyst's experience, and the total person-year resources available.

The next quality assurance activity is applied when data are keyed from the transcript into the computer. Some straightforward alpha-numeric checks are done and errors are corrected. The corrected data from that stage is fed into a computer system which builds up the corporation sample file with the sample observations. Edit checks are applied at this stage to verify arithmetic totals, logic checks, cross-field checks, etc. When errors are detected, the source document and the transcript are re-worked by the corporation data analyst.

In the Head Office organization subject matter experts review a batch of returns and financial statements from which data have been collected, passed through the four steps previously described, and accepted as clean data. These returns are selected on the basis of a statistical sample, the objectives of which vary from year to year: for example they may focus on a particular industry group or a particular tax schedule. After identification and correction, errors are analysed and summarized in a report recommending corrective actions; for example, improved instructions to analysts and correction of computer program errors.

A further quality assurance activity is what we call "final validate". Throughout the year as the data are being collected off corporation returns, some inconsistencies usually

emerge in the specifications or in the edit checks. As a result, new edit checks are added throughout the year. When the file is completed, all sample observations are then subjected to all of the edit checks introduced during the year. This sometimes reveals additional errors which have to be corrected and the tax return has to be re-selected and re-worked.

Finally, the data on the file are tabulated and summarized for several key fields. These summarized fields are compared to prior years' corporation sample files and to other sources of data; for example, the Statistics Canada data published in "Corporation Taxation Statistics" and the 100% data available on NRT's main administrative file. Large, unexplained differences are investigated. As appropriate, the file is adjusted.

5. FUTURE ISSUES

In the future, we expect to implement a number of changes which will impact the data quality. We are presently doing a feasibility study of automated data capture to replace the existing key edit data capture system. With an automated data capture system, the person who keys in the data would submit it to the computer program that applies edit checks and diagnosis errors in the data for immediate correction by that same person who had keyed in the data.

Our existing quality assurance activities are based largely on classical production planning and control techniques, and involve only limited techniques such as the selection of returns based on a probability sample. We are planning to examine the feasibility of introducing statistical methods to measure the quality of data and to identify tolerance limits that could be applied to judging the acceptability of the data.

If automated data capture and more sophisticated statistical techniques are implemented, we expect to have greater assurance that the quality of the data provided satisfies all the purposes for which it is collected.

ACKNOWLEDGEMENTS

Special thanks are due to the subject matter experts in Corporation Statistics Section — Jean Wyman, Wendy Blais and Madeleine Gadoua. They provided invaluable information on the existing quality assurance activities.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

USING ADMINISTRATIVE RECORD DATA TO EVALUATE THE QUALITY OF SURVEY ESTIMATES

JEFFREY C. MOORE and KENT H. MARQUIS[1]

ABSTRACT

The Survey of Income and Program Participation (SIPP) is a new Census Bureau panel survey designed to provide data on the economic situation of persons and families in the United States. Each SIPP household is interviewed eight times - every four months - over the two-and-one-half year life of the panel.

The basic datum of SIPP is monthly income, which is reported for each month of the four-month reference period preceding the interview month. The SIPP Record Check Study uses administrative record data to estimate the quality of SIPP estimates for a variety of income sources and transfer programs. The project uses statistical matching techniques to identify SIPP sample persons in four states who are on record as having received payments from any of nine state or Federal programs, and then compares survey-reported dates and amounts of payments with official record values. The paper describes basic considerations in designing the project and presents some early findings.

1. INTRODUCTION

This paper addresses issues concerning the use of records to evaluate the quality of survey estimates and describes a specific application to the Survey of Income and Program Participation (SIPP) in the United States.

Matching administrative records to survey observations on a case-by-case basis, which we call a "record check," provides useful information to survey users and designers. A record check enables the analyst to make a full range of measurement error parameter estimates for evaluation purposes. These estimates, in turn, facilitate two basic kinds of activities:

1. adjusting subject-matter estimates such as means, proportions, correlation coefficients, and multivariate regression coefficients to correct for the measurement errors; and
2. deriving more efficient survey designs that directly address, for example, the trade-offs between measurement quality and costs.

[1] Jeffrey C. Moore and Kent H. Marquis, Center for Survey Methods Research, U.S. Bureau of the Census, Room 2737 FB 3, Washington, DC 20233, U.S.A.

1.1. Basic Terms

Our focus here will be on measurement or response errors, although the record check method can be extended to evaluate other nonsampling and sampling errors also. This is not a technical exposition, but we do need to define some of our basic terms first. We assume that the survey observation from sample element i can be expressed as the sum of the true value and an error, e: $Survey_i = True_i + e_i$.

The average bias in a set of N survey observations, which we call the response bias or survey bias, is $\bar{e} = \Sigma_i / N$ and the response error variance is just Var e.

Similarly the measurement model for the administrative record observation is: $Record_i = True_i + u_i$, so that record bias is $\bar{u}$ and record error variance is Var u.

1.2. Comparison of Evaluation Approaches

The capabilities of the record check approach can be contrasted to other methods of evaluation such as reinterviews and experiments. Reinterviews and other repeated measures designs aim at estimating a very limited set of measurement error parameters, usually something called the simple response variance or the response error variance. These approaches implicitly make strong assumptions about true change over time and about either the true score or bias parameter (Marquis (1986)).

One frequently attempted remedy is to create a criterion measurement as part of the reinterview program, for example by reconciling discrepant answers with a knowledgeable respondent or by asking much more detailed and specific questions during the reinterview. But the validity of these criterion measures is suspect. Both Bailar (1968) and Koons (1973) have shown, for example, that reconciled reinterview responses are biased. And while detailed, specific questioning is often preferred to a more global approach, there is no independent evidence that it reduces measurement biases to zero - or at all. Record checks potentially provide higher quality criterion information requiring much weaker (and perhaps more realistic) assumptions for purposes of estimating survey data quality.

A different method of evaluating aspects of surveys is the experiment, such as a fully-crossed factorial design or an interpenetrated design for assigning interviewers. Analysts compare experimental groups with respect to statistics such as subject-matter means or proportions and draw conclusions about which treatment produces more or less reporting of the subject-matter of interest. What is controversial, however, is determining which treatment is "better" in a measurement sense, a difficulty that is much reduced when criterion data are available, such as administrative records.

Without criterion data, it is often necessary for the analyst to resort to strong assumptions about measurement errors such as:

- more reporting is better reporting;
- forgetting of meaningful material increases with the passage of time;
- unbounded interviews contain overreports, bounded interviews don't;
- reporting performance decays with length of interview or time-in-sample;
- people tend to be lazy and devious - they will lie to avoid being asked a detailed set of questions; and
- self reports are better than proxy reports.

Indeed, these assumptions have become part of the folklore of survey design in the western world. And yet, it is difficult to find any support for any of these assumptions from appropriately designed record checks. Experiments and related arrangements are

excellent approaches to pinpointing the sources of variation, and to untangling estimation problems of collinearity, but are often unnecessary and seldom sufficient for evaluating an existing measurement process.

In sum, these other evaluation approaches are forced to make strong assumptions about: (1) the independence of the original and evaluation measures when they are clearly dependent; (2) the relationship of the original measure to a criterion when no objective, external link exists; and/or (3) cognitive processes not supported by research.

Record checks also employ assumptions in evaluating measurements. For example, the usual way of estimating the response bias is to assume no record bias ($\bar{u} = 0$) and take the average of the differences between the matched survey and record observed values: Estimated Survey Bias $=\Sigma(S_i - R_i) / N$. While one cannot directly support the no record bias assumption, one can conduct meaningful sensitivity tests of the effects of possible violations of the assumption on evaluation conclusions. (At a later date the SIPP Record Check Study will employ these tests and other analyses to examine errors in the records.)

1.3. Issues in Designing Record Checks

Several issues merit consideration in designing a record check to evaluate survey measurement. We comment on some of the main issues here: incomplete observation designs, matching errors, record errors, true score differences, and absence of repeated measures or experimental design features.

1.3.1. Incomplete Observation Designs

Past record checks have often used one-directional or partial designs for data collection, such as when we survey people about owning library cards and check the records for those who claim to have one, or when we sample from a list of people with a diagnosed chronic disease and survey them to see if they report it in a survey questionnaire. Because these partial designs do not observe the full range of response errors in the correct proportions, they yield biased estimates of such classical measurement error parameters as the response bias and the response error variance. One-directional designs can fail to detect some or all of the true survey bias, can cause the analyst to interpret up to one-half of the response error variance as response bias, and can predetermine the sign of the estimated response bias if the measured variable is binary (Marquis (1978)). Full designs are a necessary (albeit not sufficient) condition for obtaining unbiased estimates of the desired response errors.

1.3.2. Matching Errors

The essence of the record check is a one-to-one matching of survey and record observations. This is difficult to do correctly, and matching errors (false matches, false nonmatches) will potentially bias the measurement error estimates of interest. Neter et al. (1965) show that when there are no unmatched cases, the mismatches will bias the estimates of response error variance upward. In terms of the reliability of a dichotomous measure (which is a function of the response error variance), the estimate will be attenuated by exactly the match error rate (Marquis et al. (1986)). It is therefore desirable to keep match errors to a minimum and to know something about the errors that remain.

1.3.3. Administrative Record Errors

As noted earlier, one usually has confidence that the records in a record check study are very good measures of the trait of interest. If the implied assumptions about record measurement bias and record measurement error variance are violated, this can cause the

response error estimates to be biased away from zero. For example, bias in the record observations can appear as bias in the survey observations but with the opposite sign. Feather (1972) describes this effect in a record check of physician visits in Saskatchewan, in which an apparently large survey overreporting rate was due to the record's recording a complete treatment procedure rather than the individual visits for the diagnosis. Similarly, the presence of measurement error variance in the record can cause inflated estimates of response error variance in the survey (Marquis (1978)).

1.3.4. True Score Differences

Problems arise when the survey and record systems use different definitions. This is often the case in "aggregate comparisons" of population parameter estimates made separately by each source. A common difference is in the scope of the populations covered, such as when the survey frame is limited to the civilian, noninstitutionalized population and the record includes everybody. Case-by-case matching can minimize the threats posed by differential coverage, but even estimates derived from these studies can still be plagued by differences in the concepts or the attributes of the concept. For example, our administrative records often contain the date a check was written for a transfer payment and SIPP survey respondents tell us when they received the payment. Such differences can threaten our time-related estimates of such things as telescoping response errors.

1.3.5. Absence of Experiments and Reinterviews

Evaluation record checks can detect errors but are not good at evaluating the remedies for the errors. To know how well a different survey design might perform, one must usually either test the alternative design options or arrange to estimate parameters of an underlying model from which survey designs can be derived (e.g., a model of forgetting effects). For example, an evaluation record check design can estimate and compare response errors for self and proxy respondents. Without heroic assumptions it cannot, however, suggest how the measurement error parameters would change if the survey's respondent rule were changed (say, to allow only self-response).

Similarly, a record check without a reinterview or another set of independent measures is limited in the number of basic error parameters it can estimate. For example, our initial definitions mentioned three parameters: true score, survey error, and record error. Without a reinterview (or other independent measure) there are only two measures with which to estimate the three unknowns. An additional measure such as a reinterview can help identify the estimates of the parameters in the model.

2. CHARACTERISTICS OF SIPP

Here we briefly describe features of SIPP as a prelude to discussing the record check evaluation design.

2.1. Overview of SIPP Contents

The purpose of SIPP is to provide improved information on the economic situation of people and households in the United States. It collects comprehensive longitudinal data on cash and noncash income, eligibility for and participation in Government transfer programs, assets and liabilities, labor force participation, and a host of related topics. SIPP data assist the evaluation of the cost and effectiveness of current Federal Government programs, of the potential impacts of proposed program changes, and of the actual impacts of changes when implemented. In general, the Census Bureau and other

Government agencies which have fostered and supported the development of SIPP expect it to be an invaluable tool for domestic policy planning (Nelson et al. (1985)).

Core SIPP questions - repeated in each wave of interviewing - cover labor force participation and amounts and types of income received, including transfer payments and noncash benefits from various programs for each month of the reference period. The core questions cover nearly 50 sources of income, including Government transfer payments from retirement, disability and unemployment benefits, and welfare programs such as Aid to Families with Dependent Children. Information is also gathered on noncash programs such as food stamps, Medicare, and Medicaid; private transfers such as pensions from employers, alimony, and child support; ownership of assets that produce income, such as interest, dividends, rent, and royalties; and on miscellaneous sources of income, such as estates.

2.2. SIPP Data Collection Design

SIPP started in October 1983 with a sample of approximately 25,000 designated housing units (the "1984 Panel") selected to represent the noninstitutional population of the United States. In February 1985 a new and slightly smaller panel was introduced. Additional panels are to be introduced each February throughout the life of the survey. Due to budget reductions, the sample size for new panels is currently about 15,000 households.

Each sample household is interviewed by personal visit once every four months for 2-½ years, resulting in a total of eight interviews per household. The reference period for each interview is the four months preceding the interview month. At each visit to the household, each person fifteen years of age or older is asked to provide information about himself/herself. Proxy reporting is permitted for household members not available at the time of the visit. Information concerning proxy response situations is recorded and is available for analytical purposes.

To facilitate field operations, each sample panel is divided into four subsamples ("rotation groups") of approximately equal size, one of which is interviewed each month. Thus, one "wave" or cycle of interviewing is conducted over a period of four months for each panel. This design produces steady field and processing workloads, but it also means that each rotation group uses a different four month reference period.

Beginning with the second wave of interviewing in the 1984 panel, SIPP includes reinterviews with a small sample of households about a subset of items (including program participation). These data are used primarily to check for interviewer falsifications, but may also be of some use in estimating response inconsistencies.

3. RECORD CHECK DESIGN

The purpose of the record check is to provide an evaluation of some of the data gathered in SIPP. We highlight important features of the design of the record check next, covering the samples, the administrative records, the matching approach, and the analysis.

3.1. Record Check Samples

The SIPP record check uses a "full" rather than a one-directional design; that is, the records we have allow us to validate all observed values in the survey. Design options we did not choose include: (1) checking records only for people who claimed to be participating in a program, or (2) drawing a sample of known recipients and interviewing them to determine how truthfully they report. Both of the latter designs are incomplete and will result in biased estimates of the response error parameters.

The Record Check Study restricts attention to a subset of available SIPP data from the 1984 Panel. First, the sample of people is restricted to households in four target states: Florida, New York, Pennsylvania, and Wisconsin. In the 1984 Panel this translates to approximately 5,000 households. Second, the study's sample of calendar time periods includes only the first two waves of the 1984 Panel. Figure 1 illustrates the wave, rotation group, interview month, and reference period structure for the target survey data.

Figure 1:
Survey Structure for Data Included in the
SIPP Record Check Study

Wave	Rotation Group	Interview Month	Reference Period Months										
			Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
1	1	Oct 83	X	X	X	X							
	2	Nov 83		X	X	X	X						
	3	Dec 83			X	X	X	X					
	4	Jan 84				X	X	X	X				
2	1	Feb 84					X	X	X	X			
	2	Mar 84						X	X	X	X		
	3	Apr 84							X	X	X	X	
3	4	May 84								X	X	X	X

Third, the SIPP Record Check Study focuses on the quality of recipiency and amount reporting for selected Government transfer programs. We will compare survey reports and administrative records for five Federally administered programs (Federal Civil Service Retirement, Pell Grants, Social Security (OASDI), Supplemental Security Income (SSI), and Veterans' Compensation and Pensions) and four state-administered programs (Aid to Families with Dependent Children (AFDC), food stamps, unemployment compensation, and worker's compensation).

We limited the study to four states - Florida, New York, Pennsylvania, and Wisconsin - in order to keep the study to manageable proportions. Major criteria used to select these states were: (1) the presence of a computerized, accessible, and complete record system for all target programs; (2) a large SIPP sample; (3) reasonable geographic diversity; and (4) a willingness to share individual-level data for purposes of this research. Thus, the states were selected purposively; no attempt was made to sample states to be representative of the Nation.

We requested from each participating state agency identifying and receipt information for all persons who received income from the target program at any time from May 1983 through June 1984. The identical request was made of the participating Federal agencies, with the exception that only recipients residing in one of the four selected states were to be included in the data extract.

We obtained these administrative records with the understanding that they would be accorded the same confidentiality protection as data gathered by the Census Bureau under Title 13 of the U.S. Code. Thus, the records may be used only by sworn Census Bureau employees engaged in and for the purposes of the Record Check Study. Except in the form of nonindividually-identifiable statistical summary data, the records may not be released or disclosed to any others for any purpose.

Some agencies elected to follow a two-step procedure, initially providing only recipient identifying data, with no (or only minimal) data on program benefit receipt history. Following the matching of the recipient and SIPP files the project will send back

to the agency a list of case identifiers for matched persons (plus a sufficient number of nonmatched case identifiers to assure the confidentiality of the SIPP sample). The agency will extract and return to the Census Bureau payment history data for these cases.

As noted earlier, errors in the records can cause problems for record check evaluation studies. Although several of the administrative record files obtained for this project contain very minor deficiencies (for example: not listing a middle initial; no sex designation; age, rather than date of birth; etc.), only three appear at all likely to pose major analytical problems. Two are known to be incomplete in their coverage of recipients: the New York worker's compensation file, and the Veterans' Compensation and Pensions file covering all four states. The former excludes an unknown number of cases which were "closed" (i.e., cases which had already been adjudicated and for which payments by a private insurance carrier had already begun) at the time the data base was created several years ago. The latter excludes the approximately one percent of all recipients whose benefits were sent to a financial or other institution. There are no known coverage problems with any other files. The third problematic file has complete coverage but lacks recipient address information, which can be very useful for matching.

An unavoidable problem which afflicts all of the administrative files to some extent is the discrepancy between payout date and receipt of payment; obviously, the SIPP respondent reports the latter and has no knowledge of the former, and the reverse is true for the program records. Where the payout date is close to the end of a month it may be difficult to distinguish a forward telescoping error from a legitimate difference between month of payment and month of receipt. Where there are definitional discrepancies, such as this payment date issue, our analyses will attempt to model them explicitly.

4. MATCHING

4.1. Introduction

The quality of matching has an important effect on some of the most critical response error estimates such as the response error variance. Ideally, variables used to match survey and record observations are measured without error and are able to identify an individual uniquely. The ideal, of course, is never realized.

However, the variables we have available to match surveys and records should go a long way toward minimizing the match errors. Some, such as social security number (SSN), uniquely identify an individual even if other information such as address is outdated, garbled, or obliterated or missing. For purposes not directly related to this study (although certainly of benefit to it), the Census Bureau has taken special measures to ensure that SSN information as reported to the SIPP is complete and valid. For all Wave 1 and 2 sample persons, reported SSN's and reports of not having an SSN were verified and, if necessary, corrected, by the Social Security Administration. Sater (1986) estimates that as a result of this operation the SIPP file contains a valid SSN for about 95 percent of SIPP sample persons who have one.

The wealth of other data - last name, first name, house number, street name, apartment designation, city, zip code, sex, and date of birth - is sufficient for high quality matching even in the absence of a unique identifier such as SSN. In addition, to aid us in evaluating the impact of any remaining match errors, the Census Bureau's matcher produces an ordinal measure of the goodness of the match/nonmatch of each survey observation to its appropriate administrative record counterpart.

4.2. The Census Bureau's Computerized Match Procedures

The Record Check Study uses computerized statistical matching procedures applying the theoretical work of Fellegi and Sunter (1969). These procedures were developed at the Census Bureau, primarily for purposes of census undercount estimation.

Computerized statistical matching is the process of examining two computer files and locating pairs of records - one from each file - that agree (not necessarily exactly) on some combination of variables. The process involves multiple discrete steps, but basically there are four: standardizing the common data fields in the two files which the matcher will examine to determine whether a pair of records is a match or not; sorting the two files into small subsets of records (or "blocks") which constitute a feasible number of pairs to be examined by the matcher; determining and quantifying the usefulness of each data field to be considered in the match for identifying true matched pairs; and implementing the computer algorithms which perform the actual record matching.

4.2.1. Standardization

We will process all data files in the Record Check Study - both the SIPP files and the administrative record files - through an address standardizer which standardizes the format of various components of an address (e.g., street name, type, and direction; city name; state abbreviation; etc.) and parses each component into a fixed data field. Several programs have been developed for this purpose; we currently use the ZIPSTAN standardizer developed at the Census Bureau, but may soon switch to a new generation product developed by our Geography Division.

In addition to the standardization procedures which apply to all data files, many of the files require modifications to individual data fields to ensure a common format across files for matching. Common examples of variables which pose problems of this type are sex (which can be represented by either an alpha ("m" or "f") or a numeric ("1" or "2") code); date of birth (which has many variants - e.g., "mm-dd-yy," or "cc-yy-mm-dd," or the Julian format); and name (which may be a single field or which may have separate fields for each component). Currently we prepare custom-made programs to carry out this type of standardization but a new version of the Census Bureau's Generalized Data Standardizer (GENSTAN) may soon take over this task.

4.2.2. Blocking

Blocking - establishing subsets of records for the matcher to examine in searching for matched pairs of records (e.g., Jaro (1985)) - is a necessary strategy when matching files with large numbers of records. Obviously, the probability of finding all true matches would be highest if, for each record on one file, the entire other file were searched for a match. However, for large files such unrestricted searches for matched records is simply not feasible. Blocking each file into subsets of records makes matching large files feasible, but at the cost of excluding some records from the search, thus increasing the likelihood that some true matches will be missed. Ideal blocking components, therefore, have sufficient variation to ensure the partitioning of the files into many (and therefore smaller) blocks, and are effective match discriminators - that is, nearly always agree in true match record pairs and nearly always disagree in true nonmatch record pairs. (The latter also implies that an ideal blocking component must be largely error-free on both files.)

The first of these criteria - sufficient variation - is easy to achieve; the second is more problematic. The primary blocking strategy for the SIPP Record Check Study employs the first three digits of the United States Postal Service's five-digit zipcode and a four-character SOUNDEX code derived from the sample person's/recipient's last name.

The former is a sub-state geographic indicator which generally is recorded quite accurately according to Census Bureau matching experts. The latter is a widely-used algorithm for creating a standard length, standard format code from input character strings of varying lengths. The code is comprised of the first letter of the string (here, the last name), followed by a numeric code which is based on only certain letters in the remainder of the string. The advantage of such encoding for blocking purposes is that it minimizes blocking errors due to misspellings, although it cannot eliminate such errors entirely.

Because the success of the match is so sensitive to the blocking scheme, the study will use at least two and possibly three separate blocking strategies - each employing totally unrelated blocking components - for each pair of files to be matched. This will minimize the likelihood that a true match pair will escape detection as a result of blocking. These subsequent blocking arrangements will not be uniform for all matches (because of variations in the availability of some data fields or because of known problems with quality) but are likely to include some combination of sex, month of birth, day of birth, SOUNDEX code for city or street name, or partial SSN.

4.2.3. Data field match weights

With some variation, the data fields used in the matching of the SIPP and administrative record files will include house number, street name, apartment number, city, zip code, SSN, sex, date of birth, last name, and first name. Intuitively, these fields are not equivalent when it comes to determining whether a particular pair is a match or not - agreement on sex is not as indicative of a true match as is agreement on SSN, for example. Fellegi and Sunter (1969) include, in their presentation of a general theory of record linkage, discussions of weight calculations reflecting different data fields' differing discriminating powers and how these weights feed into optimal decision rules. The Census Bureau's Record Linkage Research Staff has developed programs using Newton's method for non-linear systems (see Luenberger (1984)) to solve the Fellegi-Sunter equations, and these programs are being used in the SIPP Record Check Study to compute final match weights.

4.2.4. The computer matcher

The Census Bureau is developing a computer matcher (CENMATCH) operating on IBM personal computers, on an IBM 4361 mainframe, and on other hardware, which executes the procedures of Fellegi-Sunter on a user-defined set of data fields on files sorted (blocked) according to user specifications. The user enters the initial match weights for each field, defines the type of agree/disagree comparison for each field (whether the fields must be exactly comparable in order for the matcher to treat them as agreeing, or whether only approximate comparability is necessary), identifies missing value entries and specifies how they are to be treated (included or ignored in the calculation for a composite match weight), and sets the composite weight cutoff values for matched pairs an nonmatched pairs. The user generates the appropriate COBOL program codes to conduct a match according to these specifications through GENLINK, the Census Bureau's Record Linkage Program Generator (LaPlant (1987)).

In simple terms, the matcher: (1) searches each data file for comparable blocks of records - that is, records which agree exactly on the designated blocking components; (2) counts the number of records in found blocks to ensure that neither file's block size exceeds the preset maximum; (3) computes composite weight for all possible pairs of records in the block; (4) assigns each record in the smaller block to a paired record in the larger block according to a formula which maximizes the total composite weight for all pairs in the block; (5) applies the Fellegi-Sunter decision procedure to determine whether a pair is a match, a nonmatch, or requires further review; and (6) produces a "pointer" file

map to th• skipped records (i.e., records in a block on one file that is not matched with a corresponding block in the other file) and the paired records (matched /review /unmatched) in each file.

5. ANALYSIS

Our goals for the record check study are to estimate selected measurement error parameters for our samples of people, content, and times, and to assess how these errors relate both to each other and to variables that reflect survey design features. Our general plan is to use the matched data to estimate for each dichotomous participation variable:

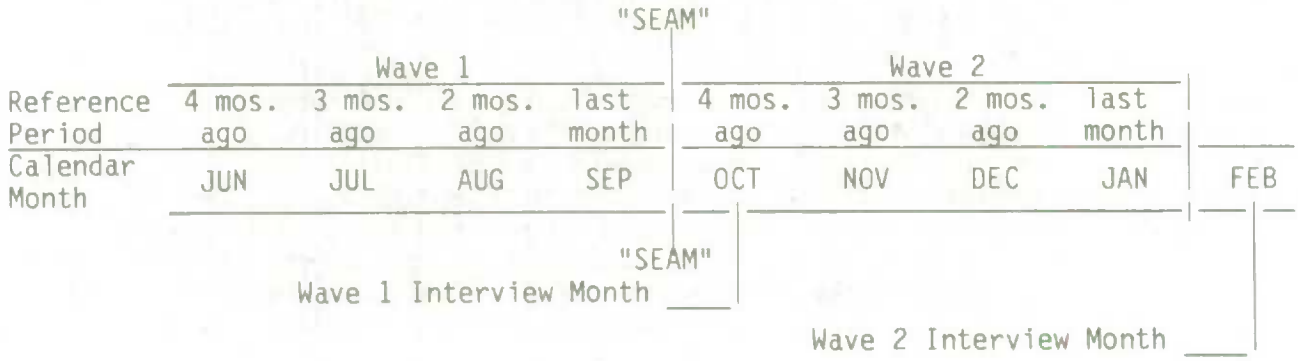
- the response bias (using the survey-minus-record difference score);
- predictors of the response bias (using logistic or probit regression techniques or possibly LISREL techniques based upon matrices containing polyserial and tetrachoric coefficients of association (Jöreskog and Sörbom (1984));
- the response error variance (e.g., derived from regression residuals);
- the conditions or groups associated with very large and very small response error variances; and
- the kinds and amounts of confusion among transfer programs that contribute to the response errors (using covariance structure analysis procedures such as LISREL).

We plan to estimate the same parameters for reports of the amounts of money received from each transfer program but we have not yet selected our basic estimation approach.

The measurement error issues to be addressed fall into one of two categories: issues which apply to all time periods and issues that require comparing errors across time periods. In the former category are estimates of the amounts of response errors for self and proxy respondents or contributed by interviewers. In the latter category are the errors arising from panel surveys with familiar labels such as telescoping, time-in-sample bias, memory decay, rotation group bias, etc. - those implying that measurement errors will differ across time periods when everything else is held constant. To this list we add what Hill (1987) has referred to as the "seam" bias in longitudinal surveys, which we discuss below.

To appreciate the applied questions we wish to address about the different time periods, consider Figure 2, which presents the interview and reference month calendar for one rotation group of SIPP respondents.

Figure 2:
SIPP Survey Time Periods for Rotation Group 1



The figure shows two interviews. The first takes place in early October and asks about what happened in September (last month), August (two months ago), July (three months ago), and June (four months ago). Similarly, the second interview taken four months later, asks about January, December, November, and October. We refer to the transition between September and October as the "seam" because it is between the reference periods covered by the two interviews.

To investigate the internal telescoping hypothesis (which asserts that events are not forgotten, just remembered as having happened closer to the present time), we will be testing whether the response bias for the early months of the reference period (June and July in Wave 1 and October and November in Wave 2) is negative and the response bias for later months (August and September or December and January) is positive, and that the two biases sum to zero.

We plan to test the bounded interview hypothesis, which says that events from the remote past are reported as happening within an unbounded reference period (June through September), but that this will not happen in reference periods bounded by a previous interview (here, October through January).

To examine the hypothesis about memory decay (that the probability of forgetting an event increases with the passage of time), we will test whether the response bias is more negative for the early months of each reference period than for later months.

The time-in-sample and rotation group hypotheses suggest that response errors will be greater in the second interview than the first, after correcting for any seasonal effects. We plan to examine this and, if we find it to be true, test some of the ideas in the literature about why it may be true. Are the sample elements that survive from the first to the second interview different, as Stasny and Fienberg (1985) suggest, or does the quality of the survivors' reporting deteriorate as the Neter and Waksberg (1966) conditioning hypothesis might predict?

We don't know yet the extent to which SIPP is experiencing these more traditional problems of longitudinal surveys. One problem for which there is evidence, however, concerns the estimation of month-to-month changes in program participation (Burkhead and Coder (1985)). Specifically, more changes in program participation take place at the "seam" between interviews (between September and October in Figure 2) than between the months covered by any one interview (e.g., between June and July or July and August or August and September). The Census Bureau has not published monthly program participation transition estimates from SIPP yet because the estimates show a pattern that appears to be affected heavily by measurement error. Moore and Kasprzyk (1984) and Hill (1987) have speculated about what kinds of response, nonresponse, or procedural errors might be producing the pattern and which set of transition estimates is more accurate. By addressing the problem with administrative data, we hope to come much closer to a definitive explanation about the role of response and nonresponse errors in producing the observed pattern.

Related, possibly, to the seam bias issue is the better-understood phenomenon that measurement error variance tends to inflate estimates of gross change or underestimate stability. Recent literature (e.g., Fuller (1986)) suggests several possible approaches to the problem. We plan to begin the empirical exploration of the measurement error effects on the transition estimates to learn whether, for example, we can base corrections for the response errors on estimates from reinterviews.

Finally, we have hinted previously at the problems that may arise in getting unbiased estimates of the errors if the records also contain errors. We plan, with the use of reinterview measures (that identify the estimate of Var e) to estimate the record error variance (Var u). However, we have no plans to relax the assumption that the records are unbiased.

6. PRELIMINARY FINDINGS

To illustrate our approach, let us look at the "seam" issue with some test data we are using to get experience with data processing procedures. Recall that the seam problem is that monthly survey reports about program participation status produce more frequent status changes between months covered by separate interviews than between other months (covered by the same interview).

Some initial questions about the survey data that administrative record information would help answer include these:

1. Are there too many transitions reported at the seam?
2. Are there too few transitions reported for other months?
3. Do the different sources report the same number of changes over the whole time period but distribute them differently?

Next we will show what we call "aggregate comparison" data relevant to these questions noting, however, that the data come from a convenience sample and do not necessarily represent any population of interest. Also note that there are a small number of cases by Government survey standards. For these reasons we will stick to descriptive statistics.

Aggregate comparisons do not involve case-by-case matching of survey and record data; in this example, however, we use exactly the same sample of 1,536 people for both the survey and record values. This eliminates differences in coverage definitions that often plague this method.

Assuming that the record data are correct, the AFDC graph (Figure 3) suggests:

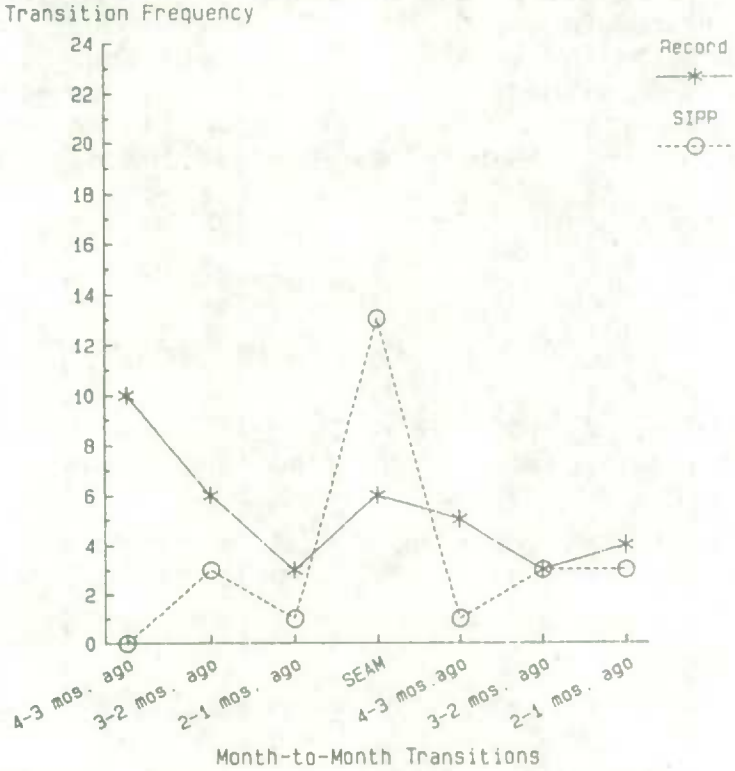
1. Too many transitions are inferred at the seam from the survey;
2. Too few transitions are inferred for the other months; and
3. Too few transitions overall are reported in the survey, a net underreporting problem as well as a time-placement problem.

Turning to the Food Stamp graph (Figure 4), we see similar but not identical trends:

1. There are still too many transitions inferred at the seam;
2. But whatever underreporting bias there is in the other months does not seem severe; and
3. Both survey and record contain about the same number of total transitions, suggesting just a time-placement problem and not a net bias phenomenon.

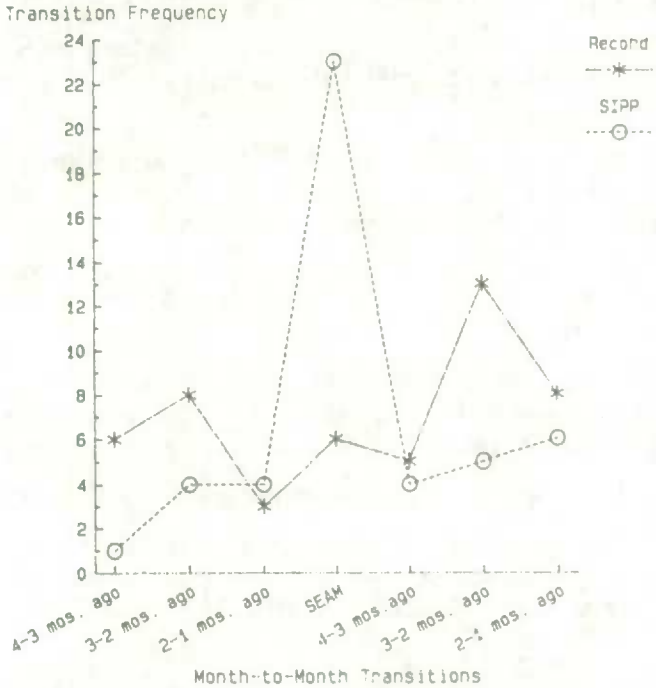
There are many more tests to be done and many hypotheses to explore before we start to draw conclusions about the nature of the measurement errors and their probable causes. We feel that the administrative record data will allow us to make important advances toward understanding the sizes and forms of these survey errors and perhaps suggest their causes.

Figure 3:
AFDC Transitions as Reported in SIPP and in Records



Record average = 5.3
SIPP average = 3.4

Figure 4:
Food Stamps Transitions as Reported in SIPP and in Records



Record average = 7.0
SIPP average = 6.7

ACKNOWLEDGEMENTS

The SIPP Record Check Study has already benefited greatly from the efforts of many people. While we cannot list here all who deserve recognition, we do gratefully acknowledge the particular contributions of: Jeannette Robinson, for preparing the multitude of administrative record files for matching; Bill LaPlant, for sharing his considerable expertise regarding the Census Bureau matcher and attendant software; Chris Dyke, for his tireless efforts to assist in making the matcher work on a new computer system; and Dan Kasprzyk, for his constant and patient support of this entire endeavor.

REFERENCES

Bailar, B. (1968). "Recent Research in Reinterview Procedures," Journal of the American Statistical Association, Vol. 63, 41-63.

Burkhead, D., and Coder, J. (1985). "Gross Changes in Income Recipiency from the Survey of Income and Program Participation," *Proceedings of the Social Statistics Section*, American Statistical Association, Washington, DC.

David, M. (1983). *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program*. New York: Social Science Research Council.

Feather, J. (1972). *A Response Record Discrepancy Study*. University of Saskatchewan, Saskatoon.

Fellegi, I., and Sunter, A. (1969). "A Theory for Record Linkage," *Journal of the American Statistical Association*, Vol. 64, 1183-1210.

Fuller, W., and Tin, C.C. (1986). "Response Error Models for Changes in Multinomial Variables," *Proceedings of the Second Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 425-441.

Hill, D. (1987). "Response Errors Around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods." Presented at the Annual Meetings of the American Statistical Association, San Francisco, CA, August 13.

Jaro, M. (1985). "Current Record Linkage Research." Presentation to the Census Advisory Committee of the American Statistical Association, U.S. Bureau of the Census, April 25, 1985.

Jöreskog, K., and Sörbom, D. (1984). *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood, Instrumental Variables, and Least Squares Methods*, Mooresville, Indiana: Scientific Software, Inc.

Koons, D. (1973). "Quality Control and Measurement of Nonsampling Error in the Health Interview Survey," *Vital and Health Statistics*, Series 2, No. 54, U.S. Public Health Service, Washington, DC.

LaPlant, W. (1987). "Maintenance Manual for the Generalized Record Linkage Program Generator (GENLINK) SRD Program Generator System." Statistical Research Division Internal Working Paper, Washington, DC: U.S. Bureau of the Census.

Luenberger, D. (1984). *Linear and Nonlinear Programming*. Reading, MA: Addison Wesley.

Marquis, K. (1986). "Discussion of 'Correlates of Reinterview Inconsistency in the Current Population Survey'." *Proceedings of the Second Annual Research Conference*, U.S. U.S. Bureau of the Census, Washington, DC, 235-240.

Marquis, K. (1978). Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations. The Rand Corporation, Santa Monica, CA. R-2319-HEW.

Marquis, K., Marquis, S., and Polich, M. (1986). "Response Bias and Reliability in Sensitive Topic Surveys," *Journal of the American Statistical Association*, Vol. 381-389.

Moore, J., and Kasprzyk, D. (1984). "Month-to-Month Recipiency Turnover in the ISDP." *Proceedings of the Survey Research Methods Section, American Statistical Association*, 726-731.

Nelson, D., Mcmillen, D., and Kasprzyk, D. (1985). "An Overview of the Survey of Income and Program Participation, Update 1." *SIPP Working Paper Series*, No. 8401, Washington, DC: U.S. Bureau of the Census.

Neter, J., Maynes, S., and Ramanathan, R. (1965). "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, Vol. 60, 1005-1027.

Neter, J., and Waksberg, J. (1966). "A Study of Response Errors in Expenditures Data from Household Interviews," *Journal of the American Statistical Association*, Vol. 59, 18-55.

Sater, D. (1986). "SSN Response Rates and Results of SSN Validation/Improvement Operation." U.S. Bureau of the Census memorandum for R. Herriot, March 11, 1986.

Stasny, E., and Fienberg S. (1985). "Some Stochastic Models for Estimating Gross Flows in the Presence of Nonrandom Nonresponse," *Proceedings of the Conference on Gross Flows in the Labor Force Statistics*, Department of Commerce and Department of Labor, Washington, DC, 25-39.

SESSION V: INVITED PAPERS

ADMINISTRATIVE RECORDS AS AN ALTERNATE DATA SOURCE

Chairperson: F. Scheuren, U.S. Internal Revenue Service

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

ADMINISTRATIVE DATA AS ALTERNATIVE SOURCES TO CENSUS DATA

J. PODOLUK[1]

ABSTRACT

In Canada various administrative data bases exist which cover large segments of the population. For example, because of universal social security programs, data bases exist covering almost the entire population aged 65 and over and children under 18. However, these data sets are complementary to the computerized income tax records which directly cover more than three quarters of the adult population. The paper discusses the scope of these records, compares tax concepts and survey concepts and outlines the kind of population data which can be developed from tax records. It suggests a possible program of evaluation of the relationship of administrative and census statistics and makes proposals as to possible improvements to tax records to enhance their statistical usefulness.

INTRODUCTION

Canada, unlike some other industrialized countries, does not have a population register. However, large administrative data files created because of fiscal and social programs have existed for decades. Increasingly, these files have covered nearly all of the population or almost complete segments of the population such as the elderly or children. However, until the advent of computers, these files were maintained manually and thus not readily accessible for statistical purposes.

When records started to be computerized in the sixties, Statistics Canada initiated early efforts to use administrative data for statistical purposes but these early efforts were not very effective for a variety of reasons. Some files were not yet computerized, the postal coding system was crude, and computers could not handle millions of records readily. Another problem was that although Statistics Canada had a mandate under the Statistics Act to be responsible for the development of the national statistical system, its rights to access administrative data files such as income tax records were ambiguous. Legislative changes established its right to access clearly.

Many of the earlier problems have now been overcome and, given the growing scope of administrative data and the growing cost of censuses, the question arises as to what extent administrative statistics could supplement or replace census data. This paper makes some assessment of this issue and suggests some further evaluations which could be carried out.

[1] Jenny R. Podoluk, Consultant, 626 Gainsborough Avenue, Ottawa, Ontario, K2A 2Y8, Canada.

1. QUALITIES OF ADMINISTRATIVE DATA FILES

There is not yet a conviction that administrative data can fully replace other sources. Part of the unwillingness to accept administrative data files as a substitute for survey and/or censuses results from the fact that they are created because of the existence of some fiscal or social program and because programs can be altered or eliminated such changes impact on the data files which are a by-product of the administrative process.

Nevertheless, there are administrative data files which are, in effect, virtually complete data files for the universe covered. It should be noted that although the quality and the coverage of some data files has improved over the last two decades, there are no significant new programs which generate large data files so that the files which can be considered for exploitation are basically the data files which already existed in the sixties when the first efforts were made to incorporate administrative data into the statistical system.

2. DESCRIPTION OF DATA FILES

The most important of these data files are the following:

2.1 Income Tax Files.

Interestingly, income tax statistics came into existence more than 60 years ago and until the World War II the then Dominion Bureau of Statistics was responsible for producing statistics. During the war, responsibility for the production of data was assumed by National Revenue. Until the advent of computers the National Revenue statistical reports were based upon a selected sample of returns for which data were transcribed in considerable detail. These samples did not carry any personal identifying information and were strictly used for statistical purposes. Taxfiler records themselves were maintained as manual records systems.

Today the personal income tax files are completely computerized and cover most of the adult population. Not only do the files contain current taxation data, longitudinal files are maintained over a number of years so that changes in the status of a taxfiler can be studied over time. National Revenue still selects a statistical sample known as the "Green Book" sample for a more detailed analysis of taxfilers and it is this sample which is used to produce the annual statistical report. Thus, National Revenue now has a series of computerized records which can be used to generate data. Aside from the transfer of records to computers the other significant developments which have enhanced the usefulness of the data were the almost universal use of the SIN number after the introduction of the Canada Pension Plan in the Sixties and the introduction of the present postal coding systems.

Canada has a number of major universal social security programs: Old Age Security, Family Allowances, Unemployment Insurance and the Canada Pension Plan. The Family Allowance and Unemployment Insurance programs date back to the war-period, OAS to the early Fifties and the CPP was initiated in the Sixties. In earlier years Family Allowance and UIC benefits were not taxable so that taxation data were not available on beneficiaries and contributors. Such income is now taxable so that tax files, for the majority of recipients, can be used to analyze the underlying population. OAS receipts are also taxable but tax coverage of this population is not as comprehensive. The tax records are the source for the maintenance of the records of contributions to the Canada Pension Plan so they encompass almost the entire working population.

Thus administrative data bases other than tax records usually only complement tax files and most have limited statistical usefulness in their own right. They are briefly summarized below.

2.2 Family Allowance Files

All families with children under 16 and with children 16 and 17 attending school receive family allowances. Thus, some data on the number of families and the number of children by age are available from these records.

2.3 Old Age Security Records

These cover virtually the entire population aged 65 and over and some widowed spouses and wives aged 60 to 64 in a low income status. Where incomes are low pensioners may qualify for a Guaranteed Income Supplement. Recipients of the latter must provide an income statement. In future, coverage of the elderly may be gradually less complete as older immigrants may not qualify for benefits.

2.4 Canada Pension Plan

The Canada Pension Plan was introduced twenty years ago. Since its inception the income tax system has been used to collect and record the contributions of the labour force to the plan. Further, CPP benefits are taxable so that tax records, as in the case of family allowances and OAS receipts, are a source of information on the characteristics of beneficiaries. Although separate files on contributors are maintained for the administration of the plan, much of the basic data is extracted from income tax records. The limitation of these files is that Quebec administers its own plan so that the contributor files only cover the labour force in the remaining nine provinces. Further, contributions to the plan have an annual earnings limit so that it is not clear whether the contributions files contain data on actual earning as compared to the amount of earnings upon which contributions are made.

2.5 Unemployment Insurance Commission Files

Initially UIC benefits were not taxable but with the considerable extension of benefits implemented some fifteen or so years ago National Revenue became an important element in the administration of the UI program. UI benefits became taxable and National Revenue also became the collector of premiums. Thus again income tax records are a source of information on both contributors and beneficiaries.

2.6 Summary

Income tax statistics are thus now the fulcrum of the national administrative data bases. The other administrative data sets can be considered to be complementary or supplementary to taxation data. In some cases, such as the CPP data on contributions, the data are derivatives of taxation data.

The point has already been made that administrative data bases may have an inherent instability because changes in programs or legislation can affect the scope and character of the data available as an input to such data systems. And also as noted above, most of the large data sets now in existence have evolved over some decades so that virtually none are of recent origin.

However, the government has indicated that tax reform has a high priority over the next few years, and in fact, some taxation restructuring could occur before the next decennial census. A series of proposals regarding changes in the personal income tax structure have been released recently by the Minister of Finance. He proposes to eliminate exemptions and to replace them with various tax credits along with rebates to non-tax payers for sales taxes. The new proposals are not likely to result in any diminution in the number of tax filers and, in fact, if new sales taxes are introduced along

with greater rebates the proportion of the population filing tax returns should continue to increase.

The next sections will discuss in more detail the special features of the administrative data files, the nature of the current utilization, the evaluations for 1986 which should be made of census versus administrative data, the further potential uses of administrative data, the limitations of and the enhancements necessary to administrative data if these are to be proxies for/or alternatives to a 1991 Census.

3. INCOME TAX RECORDS

3.1 Unit of Taxation

Unlike some other countries, such as the United States, the Canadian tax system does not tax spousal incomes jointly for married couples but rather each spouse must file an individual return to report incomes subject to tax or to claim certain credits such as the Child Tax Credit. Thus the Canadian tax system is primarily based upon individual income.

Although periodically recommendations have been made that Canada move to joint filing because of the inequities that individual filing produces and because, to an increasing extent, some benefits require a joint disclosure of income, the tax system has not been changed to allow for joint filing. Given the current feminist emphasis on treating women as individuals who should have their own rights to pensions etc. it may not be politically feasible to move to a joint filing system.

The result of the current system, however, is that the present tax and social security system is riddled with all sorts of anomalies in the treatment of spouses. Examples: A married couple with a combined income of $25,000 consisting of two incomes of $15,000 and 10,000 would pay lower taxes than a married couple with one income of $25,000 although in the latter case a marital status exemption would be available to the income earning spouse. Family allowances are payable to mothers but are taxable sources of income and must be reported by the spouse with the highest income, usually the husband. Similarly the highest income recipient must claim child care expenses even if these are shared by spouses. On the other hand, to claim GIS supplementation spouses must report their joint incomes. In common law marriages, common-law spouses can qualify for certain benefits such as pensions but a spouse is not not allowed to claim a common-law partner as a dependent upon a tax return. In other cases, where it is to the benefit of the tax filing spouses, some income can be added to their income even if the non-tax filing spouse is the original recipient, for example, of dividends.

As a result, National Revenue does not attempt to produce statistics for married couples (or families) and the production of family-oriented data requires record linkage of returns. The four main variables available for record linkage are the social insurance number of the spouse, marital status, name and address. This full range of information for linkage purposes is only available for most spouses living at the same address but even here data may be imperfect if one spouse has a different name from the partner. This could be true because a wife may choose to retain a maiden name for professional purposes or because a common law marriage is involved.[2] If spouses maintain different residences because jobs are in different locations or for other reasons the linkage of records can become more tenuous.

Income data collected on surveys and censuses at Statistics Canada also use the individual as the unit of observation but because they invariably collect data for each individual resident in a household and data on the individual relationships within a household, these data can be aggregated in flexible combinations such as the incomes of

married couples, census family income, economic family income and household family income. All of these units are defined in terms of occupants of the same household. A census family consists of a husband, wife or mother/father and unmarried children while an economic family consists of all relatives living in the same habitation. A household is defined as consisting of all occupants of the same dwelling which is a physically defined unit of residence.

Aside from data on individuals, the best additional aggregations which can be attempted from taxation statistics are husband-wife incomes. The construction of census family incomes can also be attempted but, for reasons which will be commented upon later, such data can be somewhat less successfully constructed while economic family income is a more remote possibility.

A demand exists for household income but users who prefer household to family income probably do not appreciate the differences between family and household incomes and their interpretation or the limitations of the concept of household income.

On income tax records the only basis which would exist for constructing household income would be the address of the tax filer. This might or might not produce a household income concept which would equate with the Statistics Canada concept. For example, two apartments might exist at the same address but these might not be differentiated on the tax returns so that two households under the Statistics Canada definition would be treated as one household on taxation returns thus exaggerating real household income. Some household occupants, for tax purposes, might use the address of tax discounters in filing a tax return and thus their income could not be associated with place of residence. This would result in an understatement of household income.

For most purposes, except possibly for the analysis of housing characteristics, family income is probably a superior concept to household income. Generally average household income in an area is higher than average family income but the reverse may be the case where a geographic area has a substantial proportion of households consisting of one person. One person households usually have lower incomes than family households and thus to interpret household versus family incomes, one should differentiate at least between households consisting of one person versus households of two or more persons.

The Canadian concept of the household is a variation of the concept used in the United States and is peculiar to U.S.-Canadian statistics. Other countries have a concept which is not tied to the occupancy of a particular housing unit but rather a concept which defines households as person occupying the same housing unit and pooling income and expenses. This is basically the definition employed in the Family Expenditures Survey and

[2] In Quebec legislation was passed several years ago that married women legally retained their maiden names after marriage. For Quebec administrative purposes apparently some records are being converted to maiden names, for example, health insurance cards carried by married women even where the marriage occurred years ago. At National Revenue some married women are still filing with their married names as previously while others are filing under their single names. Since married couples do not have to report the full name of spouses (only first names) this is creating a growing problem of matching spousal returns for those returns which have only come into the system recently (For returns which have been filed for five or six years older returns contain the required information). The situation will become more complicated where children exist because children can be assigned either parental name (girls use mother's name or boys use father's name) or a combined parental name. Apparently in future, parents will agree on which surnames will be assigned to the children. This suggests that representations should be made to National Revenue to request the full name of spouses and not simply the Christian names.

is probably the concept that users want rather than the actual concept represented by the Canadian definition.

In summary the focus of developing new data series should be on spousal incomes followed by census family incomes if possible. Household incomes would appear to be of limited value given the difficulties of matching Statscan concepts and given that family income is probably the most significant indicator of levels of living, any focus on household income would probably not be worth any diversion of resources at present.

3.4 Population Coverage of Taxation Data

Table 1 presents data on the percentage of the population filing tax returns in 1983 by sex and age[3]. Taxation statistics cover a higher proportion of the male population than of the female population although the non-filers among the latter may be wives with little or no earnings. The groups with the lowest coverage are those under 20. It is assumed that taxfilers under 20 are in the 15-19 age group, the majority of whom are probably students with incomes too low to require filing. The aged are the other groups with low filing, especially women 60 and over and males 70 and over. Women in all age groups have a lower labour force participation rate and elderly women are more likely to be dependent on government pension payments for all or most of their income.

Interestingly enough the proportion of women age of 20 to 29 filing is somewhat higher than the proportion of males filing although after the age of thirty a higher proportion of males file than females. This is probably due to mothers filing for the child tax credit. Even so 91% of males and 84% of females in the 20 to 64 age group are directly represented in tax files. These ratios are 83% and 74% of the total population aged 15 and over. The non-filing by the youngest age groups has the greatest impact on coverage exclusions.

The lower coverage of women is probably due to two reasons: (a) women have a lower participation rate although this has been rising while the male rate in the older age groups is falling (b) older women are more heavily reliant on government pensions which are sources of income which may be entirely or partially non-taxable.

The annual work patterns data from the monthly Labour Force Survey also provide some indications of the population missing from the tax files. The work patterns survey provides estimates of the number of persons reporting labour force participation in 1983. Table 2 compares statistics from the two sources. A comparison of the number of tax returns filed by sex and broad age groups as a percentage of the number of 1983 labour force participants in 1983 shows the following:

---

[3] Comparisons in the following sections are made using Green Book statistics. Given the way the sample was selected similar comparisons with 100% data may produce somewhat different results, Totals for 1983 were not available from the master files but it would be useful to carry out coverage comparisons using 100% data. It should also be noted that many persons not filing tax returns are dependents of filers so that it is estimated that directly and indirectly over 90 percent of the population is represented on tax returns.

Table 1

Percent of Population Filing Tax Returns by Sex and Age, 1983(*)

Age	Males	Females	Total
	(per cent)		
Under 20	35.3	31.6	33.5
20-24	86.7	89.1	87.9
25-29	93.0	93.7	93.3
30-34	92.9	92.6	92.7
35-39	92.3	90.0	91.6
40-44	91.3	88.1	89.7
45-59	92.9	81.6	87.3
50-54	92.1	73.5	82.8
55-59	93.3	63.0	77.8
60-64	89.8	56.4	72.1
65-69	86.2	53.0	68.2
70+	69.0	45.5	55.1
Total	83.4	73.7	78.5
Total age 20+	89.3	78.0	83.4
Total 20-64	91.3	83.8	87.5

(*) Number of tax filers as a percentage of the population by age at Dec. 1983 estimated by interpolating 1983-84 population estimates by age. Data are from the 1985 Taxation Statistics which are based upon a 2.9% sample. Coverage may be higher on the 100% files. A question exists as to whether the data include returns filed on behalf of persons deceased in 1983 or early 1984. Some evidence to this effect exists. If so, coverage of the older age groups may be overstated.

Table 2

Tax Returns Filed as a Percent of Labour Force Participants by Sex, 1983

Age	Males	Females
15-24	64.7	63.8
25-44	94.8	92.5
45 and over	89.8	63.9
Total	85.9	75.1

As has been indicated in the previous discussion certain social security payments are now subject to taxation. Income tax data can be compared with administrative data to calculate what proportion of recipients file and what proportion of receipts of such benefits is reported by them on an income tax returns. The three types of social benefits subject to taxation are Family Allowances (including those paid by Quebec), Old Age Security pensions (but not the Guaranteed Income Supplement), Unemployment Insurance Benefits and Canada and Quebec Pension.

In December of 1983, some 3,634,811 families were in receipt of Family Allowances. Total amounts paid in 1983, including payments by Quebec, amounted to $2,487 million. The number of tax filers reporting Family Allowances on 1983 tax returns was, in fact higher than the above number, some 3,722,799 tax filers in total. The discrepancy would be due to the fact that some tax filers in receipt of family allowances during 1983 would no longer be in receipt of such allowances by December. Because of the decline in the birth rate the number of families receiving allowances over time has been falling. On the

other hand, the coverage of benefits received is 90% of total payments. This means that 10% of benefits were not reported on tax returns. This is a rather surprising total because it is to the advantage of low income recipients of family allowances to file tax returns to receive the Child Tax Credit.[4]

Because of the lower taxfiler coverage of the elderly less of the Old Age Security receipts are likely to be reported on tax returns. In 1983 the number of persons reporting OAS pensions was 59 per cent of the number of person in receipt of such pensions in Dec. 1983 and the amounts reported were 58 per cent of aggregate pensions paid.[5]

There are no unduplicated counts available of the approximate numbers of recipients of Unemployment Insurance benefits or Canada and Quebec Pension Plan benefits in 1983. However, it is possible to calculate the proportion of such benefits reported on income tax returns. In the case of UI benefits, 94.3 per cent of such benefits were reported, again suggesting that only a small proportion of the labour force does not file a tax return. In the case of CPP/QPP payments 81.3 per cent were reported on tax returns. This suggests that, over time, as the degree of dependence of the older population on the OAS/GIS diminishes because of the growing importance of the CPP/QPP, tax coverage of the older population will improve.

3.5 Summary

Income tax records have a very high coverage of the income receiving population aged 20 to 64. They are less satisfactory for the older population although the coverage is likely to improve over time while they have the lowest coverage of the young age groups. Some of the latter may have no income but, given the high participation rates reported on the Labour Force Survey, the likelihood is that they are in receipt of incomes which are too low to be taxable. One of the areas for possible evaluation for the 1986 Census data is to identify the possible characteristics of non-tax filers in the younger age groups. Family Allowance and OAS data in conjunction with taxation data, in the absence of a census, could perhaps fill some of the gaps in taxation data and this has to be considered further.

4. CONCEPTUAL RELATIONSHIPS to STATISTICS CANADA DATA

The basic data common to taxation records and censuses and surveys are age, sex and marital status of the individual.[6] To a somewhat lesser extent other common data are income data which are, in some respects, conceptually different as well as less complete and precise data on place of residence. Tax statistics provide limited data on occupational classifications and no data whatever on attributes such as immigration

---

[4] Comparisons of family allowances reported on the 100% files apparently show 98% being reported. This calculation should be checked to see whether Quebec payments were included in the comparisons. Federal payments alone were 97% of the Green Book figures in 1983 but the correct comparisons are federal + Quebec payments compared to National Revenue figures.

[5] Approximately $100 million of OAS payments appear to have been made to persons living abroad and about $34 million of CPP/QPP payments also seemed to go abroad. The comparisons above are made after adjustments for payments which appear to flow to non-residents.

[6] The continuing sample surveys such as the Labour Force Survey and the Census have standardized concepts so that comparisons of Taxation concepts and Statscan concepts apply to both Census and surveys.

status, place of birth, ethnicity or educational attainment. These aspects will be discussed in more detail below.

4.1 Income Concepts

The Census and the Surveys of Consumer Finances measure money income adhering, with some exceptions, to the money income concepts of the Personal Income sector of the National Accounts. This concept excludes as income inheritances, capital gains, windfall receipts such as from insurance policies and lotteries, and gambling gains and losses.

Some of the incomes measured in Statistics Canada series are not taxable, some sources are measured artificially for tax calculations while some receipts not considered as income for Statscan purposes are taxed by National Revenue. These are the items affected.

a) Non-taxable income - Some sources of transfer payments are not subject to tax and not reported on tax returns. The main items are: Guaranteed Income Supplement (payable to low income elderly), provincial income supplements, Veterans Allowances, Veterans Pensions and Social Assistance and Social Welfare Benefits.

b) Artificial income - Dividends receive a special treatment for tax purposes. Until recently actual dividends were not included in total income but rather an amount grossed up an additional 50%. It is this amount which is shown as an income component in the tax file so that actual income is artificially raised by a one half overstatement of dividends received. This grossing factor has now been reduced to one-third so that there may be an apparent rather than a real drop in dividend income for 1986. The reduction in grossing-up means that the overstatement of dividend income will drop to one-third. Taxation statistics could be adjusted to recalculate actual dividend income (Under the new tax reform proposals grossing up may be reduced further).

c) Taxable receipts not measured in Statscan Census and survey statistics -National Revenue taxes some income in kind or fringe benefits received as supplementation to wages and salaries. Examples are employer payments of health insurance premiums, benefits from use of a company car, free room and lodging etc. These are considered to be part of wages and salaries and these wages and salaries as defined for tax purposes include not only cash wages and salaries but also a part of what is defined to be supplementary labour income in National Accounts. It is impossible to adjust out such labour income.

A second deviation from Statscan treatment of income is the case of interest income earned on bonds. In the past, Statscan asked respondents to report actual interest received in cash.[7] National Revenue gave tax filers the option of either reporting interest cashed or interest receivable but not cashed. A substantial proportion of taxfilers used the cash method although owning bonds which accrued interest which would not be cashed in for some years. Tax policy was changed last year to make taxfilers report compounding interest at least every three years even if the interest has not been converted into cash. Over time the effect of this will be to diminish the attractiveness of compounding interest bonds and investment certificates but from now on it will be difficult to know whether interest reported on tax returns represents interest received or interest accrued but not received.[8]

[7] This was a deviation from National Accounts concepts which measure interest income accrued or earned annually.

[8] Presumably the SCF and future censuses will have to decide whether to adopt the National Revenue treatment.

Capital gains are considered to be taxable income for tax purposes although only one-half of capital gains is subject to tax. Given the current life-time exemption of $500,000 from taxation (to be reduced to $100,000), less of such income will be reported on future tax returns. Such income can be adjusted out for comparability with Statscan data which do not consider capital gains to be income.

Other income which is taxable but which may not be considered as income on household surveys are scholarships, bursaries, fellowships, energy conservation grants as well as income from illegal sources such as gambling and prostitution. The treatment of registered retirement savings plans may also lead to some duplication of income within an individual year. For example, if a contribution is made to an RRSP within a year and a withdrawal made the same year National Revenue would treat the withdrawal as income even if the withdrawal represented contributions made out of current year income which is reported on a gross basis on a tax return.

In summary, some adjustments can be made to taxation income concepts to ensure greater comparability with Statscan data but it is impossible to completely align the two series.

4.2 Geographic Classifications

The Small Area and Administrative Data Division has by now acquired very considerable experience with the geographic coding problems of tax records so that it is unnecessary to discuss them in detail. Apparently with few exceptions records now carry the postal code but this, by itself, does not eliminate the problem of matching National Revenue data on taxfilers to census data for a variety of reasons. Some returns are filed on behalf of taxpayers by income tax discounters (such as H.R. Block), lawyers, accountants or trustees so that the address available on the tax return is that of the agent filing rather than that of the taxpayer. Apparently National Revenue is taking steps to obtain home addresses of taxpayers filing through tax discounters.

National Revenue codes localities but this coding is apparently unreliable because the Department accepts the taxfilers description of locality rather than attempting its own verification. Thus if a taxfiler reports Ottawa as an address rather than Nepean or Toronto rather than Scarborough the assigned locality codes would be Ottawa and Toronto. National Revenue has also given up attempts to code to the Statscan Metropolitan areas. Thus to correct for miscoding or to match to Statscan geographic delineations it is necessary to use a postal code conversion file.

While this is feasible for clearly defined postal areas it is apparently impossible to match to Statscan geographic areas in less densely populated areas such as returns of taxfilers living on rural routes which may cross several municipal boundaries[9]. Thus urban agglomerations can be matched from the postal codes, rural areas are difficult to disaggregate.

Another future matter of concern will be the possible impact of the post office's growing use of box deliveries rather than door-to-door delivery in newly built up areas. The precision of address designations on such mail will need to be monitored closely.

4.3 Family Formation

Experimental work is already in progress to attempt to construct family income data from taxation files and the main problems have already been identified. Briefly, husband-wife returns can be matched with almost complete success but more work is necessary to

[9] For example, the village of Appleton in Lanark is 3 miles from Carleton Place but 5 or 6 miles from Almonte but for postal purposes is classified as R.R. 3, Almonte.

solve the problems of identifying common-law married couples as distinct from single parent taxfilers and persons not in families.

Married couples filing tax returns are asked to report the spouse's social insurance number. Most do so that the information available for matching consists of data on name (usually), social insurance number (usually), marital status and address (usually). Names may not match because a spouse may prefer to retain a maiden name or the name from a previous marriage[10]. Address may not be an indicator because spouses may be geographically separated or because spouses may file through agents, use business addresses etc. Work to date suggests that the availability of the social insurance number allows for successful matching of spouses in almost all cases. However it should be noted that such matching may result in matching spouses not residing in the same household and thus this would not constitute a family unit under Statscan definitions since a sine qua non is that both spouses must live in the same household as their usual residence. It might be useful in matching married couple returns to analyze whether they are identified at the same address, the same locality but different addresses or different localities.

The census family is defined as consisting of a married couple or single parent with or without never married children in the same household. Identifying children presents more problems if they file returns in their own right. If the surname and address are the same and the marital status single one can assume that they are children although the ages have to be checked against parental ages since relatives such as brothers, sisters or nephews and nieces could also be mistaken for children. Another problem exists where a child may be a step-child of one spouse and thus have a different name from that of the parent. Problems may also exist in immigrant households where surnames may be constructed differently from the traditional use of surnames in Canada.

The data on parental returns should allow for the construction of family size and the estimation of the number of children in the family although even here one can envisage problems if a child files a return in their own right but because of a low income, may still be claimed as a dependent on a parental return. Where children file returns some work may be necessary to determine whether the children are still claimed as dependents. In 1983, for example, children could be classified as dependents if their own income was below $3,870[11].

The main problems in identifying families from tax records arise in respect to differentiating between single parent families, common-law unions and persons not in families. For income tax purposes a taxfiler cannot claim a common law spouse as a spousal dependent. Thus two persons living together in a common-law union with no children present have to file as two separate individuals. A parent supporting children may use the dependency of one child to claim the equivalent of a marital status tax exemption. Thus tax returns of a parent with dependent children can be identified although even here there might be some problem of family identification as per Statscan data because the children may not be resident with the parent claiming them as dependents.

In actual fact, a household which appears to consist of a single parent situation may be a common law situation (If both common-law spouses had children of their own they might appear to be two single parent families). Research is being carried on by the Administrative Data staff to explore the feasibility of estimating common-law families as distinct from single parent families. This involves identifying all residents at the same address and, in the case of single parent families, if a person of the opposite sex whose

[10] The new Quebec problem has already been commented upon.

[11] Cross-checks are being carried out as to whether children filing tax returns may also be claimed as dependants to ensure that such children are not included twice.

age is within twelve years of the single parent also resides at the same address it is assumed that this might be a common-law marriage. Where there are no children but where two adults of the opposite sex reside at the same address this is also considered to be a possible common-law union. This has the effect of reducing the estimated number of single parent families and persons not in families. The estimates of the number of single parent families still appear to be high relative to Census statistics and the number of common-law marriages low. The estimated number of non-family persons in taxation statistics has been reduced by the attempts to estimate common-law marriages and single parent families.

A relaxation of some of the search conditions could lead to some further improvement in the estimates of single parent families especially those consisting entirely of adult children and an elderly parent. In census statistics where a parent and never married children reside together this is considered to be a family unit regardless of the ages of the parent and children. In the current work on constructing families an age restriction of 29 and under has been used as a filial age for matching a parent with a never-married child. Given the aging of the population and the growing number of very elderly parents some of the taxfilers classified as non-family persons in the current tax data research may, in fact be children residing in the same household as a parent. For example, such families may consist of an eighty year old mother and children who could be fifty or fifty-five.[12]

4.4 Relationship Between Statistical Sample and Master Files

The main statistics published by National Revenue are the annual statistics on individual returns published in "Taxation Statistics". These contain somewhat more detail than is available from the main file and are based upon a sample of tax returns. In 1983 the sample size was approximately 440,000 returns or an overall sampling ratio of 2.9 percent of returns filed. The sample is not a simple random sample but in 1983 consisted of 588 regular strata with five characteristics used for stratification: (1) Source of income (3 categories) (2) urban geographic area (cities with similar population sizes have been grouped within each region for a total of 15 urban groupings) (3) rural geographic areas (areas not included in an urban area are sampled as a single area with 12 such areas in Canada and a non-resident category) (4) tax status (taxable or non-taxable) (5) income range (4 for taxable and 3 for non-taxable). There are two additional categories -taxfilers who do not fit into these strata and taxfilers with unusually large incomes or deductions. Sampling ratios range from 2% to 100%. The sample was selected by a sequential count by means of a computer. Sampling ratios are lowest for employment income in dense population areas of the provinces of Ontario and Quebec and heavier ratios are used in sparse areas and classifications. High incomes and incomes from self-employment are more heavily sampled.

4.5 Occupation

The green book sample rather misleadingly classifies returns by "occupation". This is not a genuine occupational classification but rather a categorization of returns by principal source of income. Thus if investment income is greater than earned income the taxfiler is classified as an investor rather than as an employee or a self-employed person. Where persons are classified as employees, the categories used are where the person is employed rather than occupation per se for employees of businesses, of institutions, teachers and professors, government employees - federal, provincial, armed force, federal crown corporations, provincial crown corporations and unclassified employees. More precise classifications are attempted of the self-employed business proprietors and

[12] In future work the age restriction of 29 may be removed.

professionals where the classifications of business proprietors are by the type of business (construction, public utilities, etc.) while for self-employed professionals there is a categorization by a limited number of occupations (doctors, accountants etc.).

It is not clear what determines the classification. Some returns do not have to provide information on occupation and employer while returns of persons unemployed at the time of filing would have no employer. The questions are so vaguely formulated that there is probably no consistency in reporting. In the case of the self-employed it is not clear whether the classification is based on the T1 General answers or on the profit and loss statements of the business or profession.

In theory taxation statistics can be used to approximate the Statistics Canada concept of the gross annual labour force (exclusive of unpaid family workers) using sources of employment income reported on tax returns. However it is impossible to do so from the published data because the published statistics show the number of persons reporting each of eight categories of employment income and labour force participants may have more than one source. Thus, in 1983, the statistics show 11,196,000 persons reported wages and salaries but only 9,940,000 tax filers are classified as employees. The 1.2 million filers not classified as employees account for only 3 percent of aggregate wages and salaries reported.

5. EVALUATION OF ADMINISTRATIVE DATA AS A SOURCE OF SOCIO-DEMOGRAPHIC DATA

The next sections discuss what kind of data could be developed from administrative records which could replace census data if there were no census in 1991 or if the Census was restricted to something like a head count census.

As has already been noted some of the concepts in administrative files differ somewhat from the statistical concepts of Statscan so that administrative data might be similar but not completely comparable. Listed below are the annual data series which could be generated from administrative data files as they now exist. A later section will comment upon changes which would be required on tax returns to enhance the statistical potential of tax data.

Taxation data as they now exist supplemented by other data such as OAS records and Family Allowance records could be used to generate the following data series by geographic areas:

5.1 Population estimates by age, sex and marital status.

Currently Statscan produces population estimates annually by sex, age and province. These are benchmarked with the census and inter-censally are estimated from vital statistics and migration data derived from a variety of sources such as social security records and taxation data.

The Census data also provide similar data for sub-provincial regions. Taxation data either directly or indirectly provide statistics on the population by sex and, in the majority of cases, age. As is evident from the coverage calculated earlier the main problems are data on the younger and older population. Estimates of the number of young dependents can be derived from tax records from parental claims for tax exemptions for wholly dependent children. The amount of exemption claimed is a broad indicator of the age of the dependent child. What is missing is the sex and age of the child. Such data are reported on the income tax returns but do not appear to be computer captured.[13] Family

[13] In future National Revenue will have the ages (but not the sex) of dependent children on the computer records.

Allowance records can supplement tax records on the number of children by region and age but apparently also do not capture the information on the sex of the children. However, if census data are used as benchmarks the income tax and Family Allowance records could be used to project estimates inter-censally for sub-provincial areas.

The other missing groups are the elderly. Where spouses file tax returns the age of the spouse can be estimated from tax data. Where no tax returns are filed but where a recipient receives a combined OAS-GIS the age of the spouse can be determined. It is primarily the OAS-GIS population which does not file returns and the majority of these are women who are unattached individuals. The advent of the CPP-QPP has meant that most men retiring have incomes which are increasingly subject to taxation. In 1983, 88 percent of male taxfilers reported CPP/QPP incomes but this ratio was only 59 percent for women. Over time the proportion of the elderly filing will increase.

Of the total OAS-GIS population in June of 1985 men accounted for only 36% of the group while somewhat over two-thirds were married. On the other hand women constituted nearly 64 percent of all OAS-GIS recipients and, in contrast, about two-thirds are classified as single (presumably this includes the widowed population). Thus the majority of the elderly not filing tax returns are female elderly not in families and where the OAS-GIS recipients are married, ages of spouses are available.

At one time a deficiency of the income tax data was the absence of data on the population dependent on social assistance and family welfare benefits. The high coverage of family allowance payments and the large number of women filing who have little or no taxable income to report suggests that the majority of mothers are filing to receive the Child Tax Credit which is primarily paid to low income parents and all recipients of welfare with dependent children should be eligible.

Several qualifications do have to be made. The availability of the postal code makes it possible to aggregate population estimates derived from administrative data for urban areas but such data are more difficult to estimate for rural areas. Within urban areas because mailing addresses may differ from residential addresses the data for smaller areas such as census tracts cannot be estimated with the same reliability as for larger agglomerations. The problems that exist with tax records apply to other administrative files. For example 18 percent of persons receiving only the OAS had their cheques mailed to trustees, banks or post-office boxes.

Any changes within the next few years to universal social security programs may diminish the value of the Family Allowance and OAS files for supplementing the tax records for the non-filing population. However, this is likely to be counterbalanced by an increase in the proportions filing tax returns.

5.2 Migration Estimates

The Census collects data on intercensal migration, that is, place of residence at the time of the Census and place of residence at the previous census. It categorizes migration as to whether there was no dwelling change, a move within the same municipality, a move from outside Canada and a move between municipalities.

The tax records have two pieces of migration information annually-province of residence in December and filing address when tax returns are filed. The advantage over the Census are that the data are available annually and thus data are available on inter-censal migration and the Census data can be approximated no dwelling change, change within a municipality (use of postal code) and inter-municipal moves. The characteristics of migrants can be analyzed by other characteristics such as sex (except for children), age, income etc.

5.3 Gross Annual Labour Force, Unemployed Labour Force And Earnings

Taxation would appear to have almost complete coverage of labour force participants except the youngest workers such as students working part time. Because of deductions at source wages and salaries, especially, should be very accurately reported because employer slips must be filed. For the population over 20 or 24 thus the income tax data on sex, age and earnings of the gross annual labour force are probably equal to or better in quality than census data because as time passes, memory biases affect reporting. The income tax concept would approximate persons reporting labour force participation in the previous year with the exception of unpaid family workers (a vanishing breed). The data would allow for calculations of annual labour force participation and a measure of gross annual unemployment rates.[14] Tax data would also provide data on the income position of taxfilers experiencing unemployment and linkage of spousal returns would allow an analysis of the relationship of spousal earnings. The data on the sources of earning would allow for an approximation of class of worker into a number of possible classifications:paid worker only, self-employed only, both categories with wages predominating, both categories with self-employment predominating.

The Census measures labour force participation at the time of the Census and, for non-labour force participants whether there was labour force participation earlier in the year or in the previous years. Thus it identifies labour force participation during the previous seventeen months, occupation and industry and class of worker. Labour force participants at the time of the Census are the "current labour force" while labour force participants during the previous calendar year are the "gross annual labour force". The Labour Force Survey measures the current labour force but also identifies the last period of labour force participation for those not in the current labour force. The annual Work Patterns Survey supplement to the LFS measures the gross annual labour force.

Both labour force concepts are used extensively in the presentation of Census labour force data. The gross annual labour force concept is the measure which provides estimates of the proportion of the population participating in the labour force over a period of time, a larger group than the current in labour force. It is also the more meaningful concept to use in conjunction with earnings data.

The unemployment data from taxation data provide unduplicated counts of the unemployed during the year and would approximate the measurement of the extent of unemployment from the Work Patterns Survey. There is no parallel concept on the Census. Tax data would understate unemployment to the extent there are non-filers and that some unemployed may not receive U.I. payments. Conversely some U.I. beneficiaries receive benefits because of illness or pregnancy and not because of unemployment.

Statscan data classify the labour force by class of worker. Basically the classification is by current status in the most important current occupation. The classification from taxation statistics would be based upon sources of employment income.

5.4 Income Distributions

Because of the differences in concepts income tax data exclude information on the incomes of some of the population dependent upon non-taxable transfer payments and some age groups such as the young. The coverage seems excellent for the population aged 20 to 64, misses the incomes of young workers and is not adequate for the population whose income originates in welfare payments. For the non-filing elderly incomes could be estimated from the OAS-GIS administrative files.

[14] Persons who are sick or on maternity leave also qualify for UIC benefits.

For the estimation of family income the low filing ratio of the young is less of a handicap because parents claiming young dependents have to report the amount of income received by such dependents. The problem here is that such data do not appear to be captured on the computer so that while data exist they are not in machine readable form. However, if the young file themselves data would be available. For the elderly non-filer (the OAS-GIS population) incomes can be estimated from other administrative data. The segment of the population for whom individual and family incomes cannot be constructed are single parent families dependent upon social assistance.

The elderly non-filing population seems to largely consist of the elderly who require GIS supplementation. Those who qualify must report the amounts and sources of income received during the previous year. The only sources excluded are certain types of non-taxable transfer payments such as veterans pensions. Wives and husbands have to provide joint income declarations. For the proportion of elderly not filing tax returns the OAS files could be the basis for estimating incomes.

Women on social assistance present special problems as they file tax returns to receive the Child Tax Credit but their main sources of income are non-taxable and not reported on tax returns. Some comments will be made on the problem in subsequent sections.

5.5 Family Statistics

The work on family formation with taxation statistics has already been commented upon. The following series are possible from taxation data: individual statistics, data on married couples, and data on a census family basis (but not for economic families). I have already summarized my reservations about attempting to generate household statistics.

All of the above data can, of course, provide cross-classifications such as income or employment income by sex, age and marital status, family income by ages of spouses or by family size etc. As is evident from the summary although income tax and other administrative data cannot replace the whole range of data collected on the 1981 and 1986 Census they can provide proxy or supplementary data for small areas on an annual basis. The geographic building blocks have to be postal codes.

6. SUGGESTED EVALUATIONS OF ADMINISTRATIVE DATA

If resources permit post censal, evaluations should be carried out on census versus administrative statistics. Possible evaluations are suggested below in three categories: Macro evaluations, micro matching and research on improvements to tax records for statistical purposes.

Macro evaluations would be the easiest to carry out as these would primarily require tabulations of data and analytic resources to carry out evaluations. Micro matching would require special resources and a budget which, even if available, might not be completed in time to influence the 1991 Census but which would be a better evaluation of the relationship between census and administrative data concepts and which would point the way to improvements in administrative data series. The research on improvements to tax records for statistical purposes would require even more resources and could not be completed before the planning of the 1991 Census is finalized but which could provide essential information on the degree to which the tax returns could be improved to replace subsequent censuses such as a 1996 Census.

The original paper contained more detailed recommendations as to evaluations and research which could be carried out to better interpret census and income tax data. The following sections are briefer summaries.

6.1 Macro Evaluations

Basic tabulations should be made at regional levels of both income tax and census statistics to evaluate income tax coverage. Adjustments to concepts should be made where possible to enhance comparability. Suggested comparisons are of series such as total income by sex, age, marital status and size of income. Calculations should be made of the percentage of the population in receipt of income by sex and age and the proportion of the population filing tax returns by sex and age.

Another set of evaluations should be made of the labour force coverage on income tax returns versus coverage on the census using similar classification.

The above tabulations are for individuals. Comparisons could also be carried out on the income of married couples by their characteristics - all married couples and married couples in the labour force.

The problem of identifying single parent families and common law marriages on tax returns has been commented upon. Research is being conducted on the reconstruction of such families from tax returns. There are greater difficulties in making macro comparisons of such families with census data and because there are more conceptual problems here, at the macro level these families should be desaggregated by family characteristics such as the age compositions and income sources for comparative purpose.

Taxation statistics have large numbers of tax filers occupationally categorized as "Unclassified tax filers" and "Unclassified Employees". The characteristics of these groups need to be analyzed further.

6.2 Micro Matching

These proposals involve actual record linkages to identify problem populations in the tax records and to carry out some studies as to whether tax record data could be supplemented by other data such as OAS-GIS data. Specifically a sample of returns of female tax filers whose returns cannot be matched to a spousal return should be linked to their census returns to analyze to what extent income tax data can be successfully used to identify single parents, common-law arrangements and unattached individuals. Further, many women with little or no income, file tax returns to claim the child tax credit. Many of these may only have income from social assistance which is not taxable. Micro matching with census records would provide data on the extent to which taxation income data are incomplete or inadequate for such returns.

The above suggestions involve sampling tax records to match with census records to investigate the problems of identifying family structures. As indicated earlier the greatest undercoverage and tax records is in the youngest and oldest age groups. A sample of census records of young persons should be matched to tax records to study the characteristics of the non-filing universe, such as relationships within the household, school attendance, labour force participants, earnings, and degree of dependency on parental income.

To evaluate the non-covered elderly population a sample of OAS-GIS records should be matched to tax records. The OAS-GIS data suggest that the non-tax filing population largely consists of women with low incomes who are unattached individuals. The non-filing population should be almost completely population receiving the GIS. The GIS receiving population has to file an income statement to qualify so that income distributions could be derived for the non-filing population.

6.3 Improvements in Tax Returns for Statistical Purposes

Tax returns are designed presently for administrative purposes which provide a statistical by-product. They do have a long run potential for providing more and better data on the occupation and industry of tax filers. This would require the cooperation of National Revenue to develop better labour market data. A number of possible approaches which can be identified are: 1) allocating more space on tax returns for questions on labour market attachment 2) have tax filers self-code occupational categories 3) use the tax files as a sampling frame to do mail surveys on labour force characteristics. Such a sample could be stratified by income levels, regions etc. There is no guarantee that these attempts would improve the labour market data and there might be tax filer resistance using tax returnes to collect purely statistical data.

7. CONCLUSION

Tax returns currently, with some supplementation can be used to provide some very important basic socio-demographic data on virtually the total population at the small area level, although rural areas are more difficult to delineate than urban areas. The data potential might be improved further if resources permitted improving the statistical potential of the tax returns with the cooperation of National Revenue. Some current problems discussed in the paper when it was written last winter will be ameliorated by charges National Revenue started introducing in the 1986 tax year. To qualify for some of the credits, tax filers had to report amounts received of non-taxable incomes such as social assistance. This would primarily affect the low income population and it would make the income reporting on tax returns more complete. This would eliminate some of the conceptual differences between taxation data and Statscan data. Further, tax reform is likely to lead to a greater proportion of the population filing tax returns. This would diminish the amount of estimation which would be necessary to develop estimates for the total population.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

THE QUALITY OF ADMINISTRATIVE DATA FROM A STATISTICAL POINT OF VIEW SOME DANISH EXPERIENCE AND CONSIDERATIONS

POUL JENSEN[1]

ABSTRACT

The paper outlines the Danish experience with the use of administrative data as a main statistical source, from different points of view: the general information value, the technical properties with the quality of the results. Furthermore it deals briefly with the problems of comparing the quality of data from different sources, and it comments on the problem of using administrative data as substitute for traditional censuses in a wider perspective.

1. INTRODUCTION

The utilization of data from administrative records as sources of statistics is not a new phenomenon, and in practice data from such sources are frequently the only possible basis for a production of statistics.

The advantages and disadvantages associated with the statistical utilization of administrative records in a traditional sense are therefore well-known. The introduction of modern technology in public administration during the past 25 years has, however, resulted in a substantial expansion of potentials for deriving statistics from data in administrative registers. This is due to the following facts:

Firstly the data in most cases are organized in computerized registers. This means that they in principle are more easily accessible and processable - and presumably of a higher quality, since computerization is necessitating systematic methods of data registration.

Secondly the establishment of administrative registers (at any rate in some countries, including Denmark) has been accompanied by the introduction of general identification systems - primarily person numbers - thus creating potentials for combining data from different sources by way of record-linking. In countries without person-numbers there has been developed methods for record-linking, which for statistical purposes have been sufficiently reliable.

Thirdly the great capacity of computers has made data collection for the administrative procedures now far more comprehensive than previously.

Fourthly in some countries registers of persons and business units etc. effectively form the framework for the collection of statistical data on a traditional basis.

Below it is attempted to outline Danish experience concerning the statistical consequences of this development.

[1] Poul Jensen, Director, General Economic Statistics, Danmarks Statistik, Sejrøgade 11 2100 København, Danmarks.

When the Act on Danmarks Statistik was passed in the Danish parliament in 1966 it was foreseen that computerization would be introduced in public administration, and the resulting statistical potentials were realized. It turned out that this development came to pass very fast. Thus value-added tax was introduced in 1967, in 1968 the Central Population Register (CPR) was set up, 1970 saw the introduction of a withholding-tax system (pay-as-you-earn), and in the first half of the 1970s a number of social reforms were implemented making use of EDP registers, in 1975 the Central Register of Enterprises and Establishments (Det centrale erhvervsregister) was created and came to be administered by Danmarks Statistik, and in 1977 a Central Register of Buildings and Dwellings (BBR) was established.

The utilization of the statistical potentials offered by these and other registers has been one of Danmarks Statistik's primary objectives right from the start. Therefore a large proportion of the general production of statistics now depends on data from administrative registers operated by public authorities, and in Denmark large-scale censuses such as business censuses and population and housing censuses have now been replaced by register-based statistics.

A similar development has occurred within the field of major statistics that were previously compiled on the basis of administrative forms. Among those statistics, which are now based on data extracts from computers, can be mentioned the external trade statistics and the unemployment statistics. Moreover some new types of statistics have appeared, e.g. concerning general sales (on the basis of VAT data) and general employment (on the basis of ATP data, i.e. data from the Labour Market's Supplementary Pension Fund). Traditional fields of statistics, such as population statistics and social statistics, have benefited from the new data potentials by substantial enlargement and improvement, so that they now offer very large quantities of detailed data. In nearly all other fields of statistics, administrative registers serve as a framework for statistics in one way or another.

There can be no doubt that the development outlined above has resulted in a large quantitative increase in the statistics production. It is more debatable whether or not the quality of the statistics is satisfactory. An exhaustive description of this very comprehensive set of problems is hardly possible in this context, but the following sections indicate some of the quality problems, which have been experienced:

- what has been the effect of the above-mentioned development on the general informative value of the statistics (section 2)
- what are the technical properties of extracts from administrative registers (section 3)
- what are the properties of statistics based on register data when measured in terms of ordinary quality concepts such as reliability, continuity, timeliness, etc. (section 4)

Sections 2-4 are based exclusively on the Danish experience with the utilization of administrative data. Section 5 briefly discusses the quality of administrative data compared with that of other statistical sources of data; and section 6 sets forth some points of view that are inspired by the international debate on the census problem.

2. THE GENERAL INFORMATIVE VALUE OF THE STATISTICS

It is an obvious fact that data resulting from administration in a certain field are well suited as primary data of statistics describing the actual activities in the field of public administration concerned. It is a different question - and has indeed been a matter of great concern - whether these data are also suited for general statistics, since their

validity depends on, and is limited by, the administrative processes in which they are created.

No doubt for certain types of data the answer is in some cases negative. However, if you look at it as a whole the matter is somewhat different.

Firstly the point of view cannot be used in relation to **basic registers** such as the Central Population Registers, the Central Registers of Enterprises and Establishments etc. whose purposes are to provide information to be used in several administrative branches and for several purposes. Such registers should not be decisively influenced by their application for specific purposes. The requirements to the contents of the basic registers must therefore be said to be to some extent in line with the requirements to the general statistics.

Secondly it is possible to combine information from different administrative systems by way of record-linking. The purpose of doing so is primarily to produce data combinations that are relevant to the various statistical fields, but an important by-product is that information obtained in this way is not necessarily dominated by the data found in one particular branch of administration. In by far the most of those Danish statistical fields which are based on administrative data there are data from more than one register, including data from at least one basic register.

Thirdly it is in some cases feasible to supplement the administrative data, when required, with data whose purpose is to increase the suitability of the information for general statistical purposes. The principal example in Denmark is the so-called workplace project, which makes it possible to carry out industrial classification on establishment basis instead of the unit (quasi-enterprise) found in the administrative register in question.

Fourthly it is more or less possible (at any case in Denmark, owing to a special provision contained in the Act on Danmarks Statistik) to influence the contents of administrative registers with a view to their statistical utilization.

The overall effect of these modifications must be said to be so strong that thesis of limited validity of administrative data does not hold as a general principle. However, two reservations must be conceded: (1) Among the many data there will still be some which is so closely interrelated with a given administrative process, that their validity may be somewhat limited if viewed in the context of general statistics. (2) The concepts used may have a different definition from those used in traditional statistical data collection. For instance, the commuting concept used in register-based statistics in Denmark deviates somewhat from the one adopted for traditional censuses of population and housing.

To what extent these reservations give qualitiative disadvantages compared with traditionally compiled statistics, can only be ascertained through a detailed analyses of the individual data and their utilization for different purposes. It is safe to say, through, that the adaptation to international definitions may be more difficult in the case of administrative data than in the case of "purely" statistical data.

Concerning the assessment of the general statistical value of the administrative data the following aspects should finally be taken into account:

- the use of general identifiers means that the same items of information can be reused in different fields of statistics. For instance is industrial classifications from the Central register of Enterprises and Establishments now used in nearly all fields of statistics, so that **consistency** is effectively achieved.
- the record-linking also have an important function as analytic tool, by making possible **cross-sectional analyses** at micro level as well as **longitudinal** analyses.

These qualities were not found in traditional statistics and are of course highly important: cross-sectional consistency improves the analytical value of the general statistics. Broad surveys based on interviewing that are alternative to cross-sectional analyses are extremely costly. Collecting data from a fixed panel of respondents merely for the purpose of setting up a data background for future longitudinal surveys is highly problematic both in terms of costs and otherwise. Retrospective data collection for longer periods of time is of doubtful quality.

3. TECHNICAL PROPERTIES OF EXTRACTS FROM ADMINISTRATIVE REGISTERS

At the beginning of the edp era it was believed by many that these data would be highly suitable as a statistical medium. It was assumed that the error tolerance of administrative data must be small both among the data suppliers and among the authorities who were to process and use the data. After all, consequences of errors may be rather awkward for the citizens and they may disturb the administrative processes. Many optimistis were of the opinion that once the systems had been developed and any running-in problems had been dealt with, the production of rapid and reliable statistics was merely a question of "pushing the button". They therefore hoped that the new potentials would make it easier for the producers of statistics to overcome some of the external problems of statistics: lack of precision and insufficient timeliness.

There were also many sceptics, but their scepticism primarily concerned the data contents and the presumed limitations to the statistical validity of the data.

As indicated in section 2, the pessimistic view has not been fulfilled. However, neither has the optimistic view. Danish experience includes a number of examples of erroneous data, of production disturbances in the administrative processes - and for that matter also in the statistical processing - even in cases where considerable endeavours have been made to overcome such problems. Moreover there are examples showing that after years of intensive work it is possible to come close to the originally optimum conditions - e.g. in the case of the basic population statistics in Denmark - but there are also examples of the opposite outcome.

It is ordinarily assumed that such problems are of a temporary non-inherent nature and that in each field they can be overcome after a running-in period, at any rate if there is no shortage of qualified EDP staff.

In view of Danish experience it is, however legitimate to ask whether the problems are of such a basic nature that they can only be overcome through systematic development of methodical counter measures.

Of course it goes beyond the scope of this paper to answer this question in depth, but Danish experience does permit to outline certain universal problems regarding the utilization of administrative register data as sources of statistical data:

1. The Communication Problems

It is a well-known fact that large-scale computer operations require efficient planning and supervision. This is especially the case when several parties are involved under a single computerized system. At the very least there are computer experts with different functions as well as users at different levels. Once this complex pattern of co-operation is extended to cover statistical utilization the process is further complicated. It then becomes imperative that the documentation should be complete and fully updated. Unfortunately last-minute changes in administrative EDP systems do occur (in statistical systems too). Different computer hosts use different forms of documentation norms, and shortcomings, errors or misunderstandings regarding the documentation do occur.

2. The Problem of Errors in Procedure

Output for statistical purposes from administrative registers are of course exposed to programming errors, procedural errors, etc. In the case of administrative output the errors usually soon appear rather evidently but this is not typically the case with statistical material. In addition, the purely statistical work is not necessarily given high priority by the administrative authorities concerned, and the statistical knowledge of their computer experts is unlikely to be particularly profound.

3. The Error Treatment Problem

In statistical processing of data there is a long tradition of error detection. A similar tradition has not yet been developed within the administrative processing of data, and when existing editing productions, is primarily directed towards the administrative use of the data.

When data are received for statistical processing they should of course be checked for errors, but at that stage the proper time for thorough error detection and correction may be past. Anyway error correction procedures across institutional borders are difficult and time consuming. Moreover the Danish register supervision authorities are reluctant to approve such measures, initiated from the statistical side. This is due to the facts, that regarding record-linking, own access etc. in the Danish register legislations are less strict for statistical registers than for administrative systems. This fact inspire the Data surveillance board to enforce the "principle of one way traffic" very strictly.

4. The Editing and Combination Problem

The processes which serve to convert primary administrative data to statistical information are very complicated. They pre-suppose careful EDP planning and supervision to an extend, that in practice may be difficult to fulfil.

5. The Lacking Possibility of Visual Control

The above-mentioned problems are exacerbated by the fact that data on an EDP medium are by nature invisible. Whereas traditional primary statistical data lend themselves to visual control, which immediately reveals the most simple errors, special procedures are required in order to make computerized data available for ordinary plausibility checking. In fact in some cases the errors are not detected until the final results are available - and sometimes not until they have been published. Obviously this is rather awkward.

The above list of problem categories shows that it is no simple matter to overcome the "technical" problems that are connected with the use of administrative register data as statistical sources.

If we look for the fundamental causes the following seem to be relevant:

a. The methods of processing used in a statistical context do not fully take into account the basic difference between traditional statistical primary material - questionnaires that are visible and systematized in advance to suit statistical requirements and administrative EDP data.

b. In administrative computerized processing it is difficult to provide exact planning, careful supervision and extensive documentation, because the implicit pre-requisite for all this - enough time - is not fulfilled, due to changes caused by external factors, for instance:

- that legislation and administrative rules are changed with short notice
- that it is necessary to change the systems from time to time to achieve greater effectiveness,
- that the technical facilities - both hardware and software - are developing rapidly and thus create their own demand for time-consuming changes.

c. The level of tolerance towards errors in the output from computerized administrative systems is not so slight as might have been expected. The many running-in problems in connection with the introduction of EDP in both the public as the private sector seem to have left the general public and the media with the impression that computer errors are a fact of life which everybody has to live with.

In view of the importance of administrative EDP data in the statistics production, producers of statistics ought to develop a strategy, aiming to minimize the effect of the disadvantages.

In Denmark such a strategy has not yet been developed, but some elements of the strategy could for instance be:

- systematic testing, at the earliest possible time in the process, of the consistency of the data with the received documentation,
- setting up procedures to replace visual control of the data contents with tabulations of key information throughout all of the processing,
- an extension of error detection procedures, particularly aimed at an early revelation of systems errors and other basic errors for instance by comparing systematically with data from the same source but for earlier periods and/or with related data from other sources.

4. THE QUALITY OF THE STATISTICAL RESULTS

As previously mentioned, the final quality of those statistics which are produced entirely or partly on the basis of administrative data depends on a number of factors, including the effectiveness of the measures taken to overcome the described problems. The following contains some general observations in this field:

1. Even though some of the problems may be of a fundamental nature the quality can often be improved once the "teething troubles" of the individual system are overcome.
2. Within the fields considered, the administrative data provide complete **coverage** and thus eliminate sampling errors. Consequently a higher level of specification can be achieved than with sample surveys and in most cases. The administrative data are not affected by any non-response problems.
3. As regards the statistical **reliability** and general properties of the administrative data, various data types must be distinguished:

   - data from basic registers should generally be reliable, at any rate the most important and the most frequently used data. As indicated, they are comparatively neutral in relation to differences in administration systems, and as a main rule they are operationally defined in a statistically expedient way.
   - data which in administrative systems are created and used as basis for decisions and calculations with direct administrative consequences command a high degree of precision, but of course that does not mean that they have no shortcomings.

For instance income data supplied by the tax authorities can of course only show incomes which have been reported and assessed.

- administarative data which are used primarily as background information may be of a low quality or incomplete.
- data collected through administrative channels, but for specific statistical purposes may also be of lower quality possibly because their priority rating is not high in the administrative context.

4. The **timeliness** of the data varies. It depends primarily on whether the administrative registers are updated continuously or merely annually. The timeliness of annually updated information may be too poor to be suitable for short-term indicator. By way of example can be mentioned that the production time of the Danish register-based labour force statistics is still around 1 1/2 years, and at best it can be reduced to around twelve months.

   In other fields timeliness is highly satisfactory. For instance, preliminary results of Denmark's external trade normally are available some three and a half weeks after the end of the month in question, and detailed statistics of the Danish population are available only two months after the end of the year.

5. It has already been mentioned that the **continuity** of the data may be affected by administrative or legislative changes, e.g. the income statistics which are dependent on developments in tax legislation. However, this problem has so far not been a major one in Denmark.

   During the transition from traditional to register - bases statistics the continuity problem was particularly severe, but naturally that is a temporary issue.

5. THE QUALITY OF ADMINISTRATIVE DATA VERSUS THE QUALITY OF DATA FROM OTHER STATISTICAL SOURCES

The previous section clearly indicates that the statistical utilization of administrative data is associated with many quality problems. In cases where alternate possibilities exist but where administrative data are nevertheless selected as the basis for the statistics, the decisive arguments usually relate to cost-efficiency and reduced responded burden.

However, an important aspect of the comparison may be overlooked, namely that other statistical sources also are associated with considerable quality problems. Thus, many statistical studies have shown that traditional censuses are subject to substantial response errors and/or errors at the stage of coding the responses at the statistical offices. Moreover, in some countries censuses are subject to problems of undercounting.

As regards sample surveys the statistical reliability depends on the sample size. Statistical theory on this possible source of errors is highly developed, and it is possible to estimate the limits of its likely effects.

However, no method can show to what extent the differences between the results of two consecutive surveys should be ascribed to the underlying trend that you wish to measure, or to the effect of statistical errors.

In addition, the non-response rate, and particularly the variations in that rate, have substantial negative repercussions on the validity of the results, and obviously it is very difficult to cope with this problem.

It would be very difficult and costly to measure more precisely the effects of the error sources inherent in the different methods of producing statistics. Direct comparisons are rarely possible, as it is unusual to produce statistics of the same contents by different methods.

In Denmark, however, we have two different types of labour force statistics: one based on the record-linking of information derived from various public administrative registers (primarily those of the tax administration), the other being a traditional telephone interview survey. The former is a complete count, and the results can be broken down into detailed geographical groups. The latter is based on a sample of some 15,000 families, and it contains information relating to around 25,000 persons aged 16-74 years.

A comparison reveals that according to the register-based survey in November 1983 the population in employment (in the age groups 16-74) numbered 2,489,000 compared with 2,458,00 according to the interview survey in the spring of 1984.

About half of this difference of 1.3 per cent is due to seasonal factors, which caused the employment in the second quarter of 1984 to be lower in the fourth quarter of 1983. This discrepancy is apparent in most of the comparable main groups, and therefore the two types of statistics largely confirm each other's results as to the overall size of employment in Denmark.

The picture is another when it comes to measuring unemployment the differences are more marked. According to the interview based labour force survey the number of unemployed in the spring of 1984 was 231,000, of whom 196,000 were registered with the public employment offices and therefore were covered by the current statistics of unemployment. However, the result of the latter statistics for the same period is 243,000 unemployed. Around half of this discrepancy is attributable to a lower response rate among the unemployed than among the employed. At any rate in Denmark this seems to be a general tendency. Thus, a follow-up survey among the non-respondents of the questionnaire based labour force survey for 1979 shows a nonresponse rate of 40 per cent among the unemployed as against 23 percent in general.

6. THE UTILIZATION OF ADMINISTRATIVE DATA AS A REPLACEMENT OF TRADITIONAL CENSUSES

The census is the statistical field where the use of administrative data as an alternative source has attracted the largest international attention.

Evidently, data from population registers can replace the information gathered through traditional censuses in countries where population registers exist. It is moreover possible in countries where a unique numbering system exists to a major or minor extent, to use "exact matching" to create information normally collected through traditional censuses by way of record-linking. In Denmark the register systems have been extended so much that the entire normal array of census data can be covered. This cannot be achieved by administrative data alone. Thus, the information about the educational achievements of the population prior to 1970 is obtained from the last traditional census of population conducted in that year. Information about the relation between the residences and the places of work of the employed population (for commuting statistics) is provided partly by Danmarks Statistik, and partly by the tax authorities through a statistical extension of a tax administration system, of the workplace project, mentioned above.

In this and in other fields Danmarks Statistik cooperates with the administrative authorities on the basis of the principle of one way traffic. This seems to be accepted by the general public.

Because the primary data contained in the registers are updated annually it would in principle be possible to conduct full-scale censuses at short intervals, but of course that is not practicable. An extract of the information is published each year as part of the current statistics in the various fields. Other information - for instance statistics for sub-districts of municipalities - is distributed through special service systems that are tailor-

made to the requirements of individual user groups, and the same information may be used for ad hoc compilations, defined by the individual users.

It has turned out that these types dissemination fulfil such a large proportion of the requirements for census-type statistics that a planned register-based census for 1986 was dropped without major reactions on the part of the users. This was done in conjuction with other reductions of the Danish statistical program, caused by severe budgetary cutbacks. In most other areas the reductions were contested by the users.

These observations may lead to questioning whether the traditional censuses are at all geared to the present-day requirements for statistics. In some countries the censuses arouse considerable antagonism in the general public, they lack timeless, and it is difficult to run the production smoothly because a new big organization must be mounted prior to each census. To this should be added that also traditional censuses have qualitative weaknesses.

A possible reply to this question is that in most countries there are no alternatives to traditional censuses, and it is just not possible to find other ways of producing statistics of the size of the population and its demographic, geographic and social structure.

But in the only other statistical field which could be compared with the census field, viz, the national accounts, the conditions are just as difficult, as it is impossible to carry out "national accounts censuses". Nevertheless, such statistics are produced by fitting together pieces from a number of sources according to complicated methods of imputation, estimation and balancing.

The question is whether something similar can be done in the census field. The problems involved in that approach are described in various works by Mr. Philip Redfern, and they have been dealt with at ISI conferences, but unfortunately not very extensively at the 1987 conference in Tokyo. It goes without saying that with its excellent opportunities for utilizing administrative data, Denmark cannot contribute much to the development in this field, but it may be worth noting that by far the largest part of the data that the Danish census statistics built on are derived from the systems of the tax administration authorities and in most countries similar systems must contain information of the same nature as census data.

If these data could be extracted, systematized and utilized for statistical purposes, most developed countries could presumably set up a basis for census-type statistics which, supplemented by sample surveys, would provide reasonably good coverage of the requirements for benchmark-type statistics of individuals. The development work concerning the integration of register data and sample survey data which is being conducted by the Central Statistical Bureau in Norway with a view to the 1990 census, could be of major importance in that respect.

It is worth pointing out in this connection that it is not necessarily a prerequisite for the linking of data from different sources that specific identification keys must exist, e.g. the person number. Other types of record-linking, "statistical matching" etc. may be useful in many cases.

It would seem therefore that it is not entirely out of the question for some countries to solve their census problems in that way, but certainly a long and tough development work is ahead, and the central statistical offices may not be able to shoulder that task alone but need support from political and administrative authorities.

7. CONCLUDING REMARKS

The utilization of public administrative data as a primary source in the statistics production yields considerable advantages, but it also entails major problems.

A major methodological development effort will be necessary, not only in countries like Denmark which have good preconditions for utilizing public administrative data, but also in countries without such preconditions, but where the regard for the response burden and for the increasing problems of achieveing satisfactory results through traditional methods all the same may make it necessary to adopt administrative data as a primary statistical source. In the author's view, these problems will be a major challenge to the central statistical offices throughout the remaining part of this century.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

EVALUATING THE EFFECT OF TAX REFORM ON CENSUS BUREAU PROGRAMS

GERALD GATES[1]

ABSTRACT

The Tax Reform Act of 1986 resulted in dramatic changes in the tax filing system in the U.S.. Many of the forms used to report income and expenses to the Internal Revenue Service are being revised; procedures are being drafted to interpret the new law and provide guidance to tax filers; and, the tax processing sytem is being overhauled to accommodate all these changes. Since the Census Bureau uses administrative data from the tax files in the economic censuses and population estimates programs and various coverage and content evaluation programs, we are directly affected by any changes to the way taxes are collected and processed. In addition, there are indirect consequences as a result of corresponding changes to recordkeeping practices of businesses and individuals who report to the Census Bureau in one of our surveys and censuses. This paper describes how the Census Bureau is preparing for life under tax reform and what steps we are taking to minimize the disruptions caused by loss of information, changes in definition, and delays in collection and processing.

The 1986 Tax Reform Act not only represented a major shift in the way U.S. citizens were taxed on their income, it also presented some major challenges and opportunities for statisticians who use these administrative tax data. What the 1986 Tax Reform Act did was to reduce overall tax rates for both individuals and businesses, shift more of the tax burden from individuals to corporations, flatten progressive tax rates but maintain current revenues by eliminating tax preferences used by higher income taxpayers, and reduce the tax burden on lower income taxpayers. It also broadened the tax base by limiting or ending various deductions, credits, or special treatments that were designed to encourage certain kinds of investment. As a result of these sweeping changes, the Internal Revenue Service (IRS) has revised many of the forms used when filing tax returns. In addition, rules and procedures are being rewritten, the processing system is being reworked to accommodate the new and revised forms, and a massive campaign is underway to educate taxpayers on how to comply with the new law.

Tax attorneys and advisors are also lining up to help clients take advantage of the new law. Financial decisions are likely to be driven by how the Act treats income and deductions. A further consideration is the phasing-in of the law. Whereas the major provisions of the Act take effect with the 1987 tax year, tax rates will decline further in 1988 and several credits or deductions will be reduced or eliminated over a period of 2 to 5 years. "Two of the more significant behavioral responses resulting from this act are the deferral of income and the acceleration of deductions. Many taxpayers, faced with a

[1] Gerald W. Gates, Program and Policy Development Office, U.S. Bureau of the Census, Washington, D.C. 20233, U.S.A.

2-year phased reduction in tax rates and the elimination — or limitation — of many deductibles, will defer income and/or accelerate deductions to minimize taxes in 1986 and 1987" (Wakefield 1987).

In this setting, the Census Bureau must evaluate how the information on data files it receives from the IRS will be affected and how to take advantage of new information. Our authority to receive these data comes from Title 26, United States Code and Title 13, United States Code authorizing IRS and Census Bureau activities, respectively. Title 26, Section 6103 (j) permits the Census Bureau access to "...such returns or return information...as...prescribed by regulation...for the purpose of, but only to the extent necessary in, the structuring of censuses and national economic accounts and conducting related statistical activities authorized by law. Similarly, Title 13, Chapter 1, Section 6(a) allows the Census Bureau to"...call upon any other department, agency, or establishment of the Federal Government...for information pertinent to the work provided for in this title." Any new application of tax data, not previously authorized by regulation, requires an application by the Secretary of Commerce (for the Census Bureau) to the Secretary of the Treasury (for the IRS) and an amendment to the Code of Federal Regulations (CFR) stating what information is needed and how it will be used.

The remainder of this paper discusses how the Census Bureau uses tax data in its economic and demographic statistics programs and how tax reform may affect those uses. An evaluation of these affects is ongoing and some initial results are provided about information that is being added, lost, potentially delayed, or processed differently. The paper also discusses taxpayer behavioral changes that may affect reporting patterns and, in turn, affect the data we get. Finally, the paper discusses the conflicts between the mission of the statistical agency and the administrative agency and how the Census Bureau is attempting to take a more active role in seeking out administrative data and ensuring its continued usefulness.

1. CENSUS BUREAU USES OF TAX DATA

Since 1954, the Census Bureau has received tax files from the IRS to support its economic and agricultural censuses and surveys. We have used these data in place of direct census inquiry for many small businesses. They have also been used, among other things, to identify business births; to update and supplement business and farm mailing lists; to evaluate the quality of census and survey reporting; and to provide a source for industry classification for certain businesses.

For the 1982 Economic Censuses, all information for approximately 1.3 million of the 3.5 million inscope single establishment firms came from tax returns. In addition, information for all of the 4.7 million nonemployers was obtained from tax files. Although these tax data account for only about 5 to 10 percent of total receipts (all multiunit and large single-unit firms are mailed census questionnaires), they represent a considerable reduction in reporting burden for the smallest businesses.

For the 1987 Economic Censuses, the next cycle of our quinquennial economic censuses program, the Census Bureau will obtain information from fourteen IRS tax forms relating to businesses (Table 1). These tax files will provide receipts data for all businesses that are not mailed a census form. We will use them to define the entire economic censuses nonemployer universe (i.e., those businesses with receipts data but no matching payroll record). We will obtain addresses for use in geographic coding and to establish the initial universe for the Surveys of Minority-Owned and Women-Owned Businesses. Finally, we will use information such as end-of-year indicator and number of months in business in the selection and control of samples for various current surveys. In all, over 75 million 1987 business tax records will be used in preparing for the 1987 Economic Censuses.

Table 1

1987 Business Income Tax Returns to be used for the 1987 Economic Censuses

IRS Form #	Title
SS-4	Application for Employer Identification Number
941	Employer's Quarterly Federal Tax Return
990	Return of Organization Exempt from Income Tax
990-PF	Return of Private Foundation
990-T	Exempt Organization Business Income Return
1040 Schedule C	Profit or Loss from Business or Profession
1040 Schedule SE	Computation of Social Security Self Employment Tax
1065	U.S. Partnership Return of Income
1065 Schedule K1	Partner's Share of Income, Credits, Deductions, etc.
1120	U.S. Corporation Income Tax Return
1120-A	U.S. Short Form Corporation Income Tax Return
1120-F	U.S. Income Tax Return of a Foreign Corporation
1120-S	U.S. Income Tax Return for a S-Corporation
1120-S Schedule K1	Shareholder's Share of Income, Credits, Deductions, etc.

In addition to these uses for our economic programs, individual income tax return information is used in our demographic programs for producing intercensal population estimates and associated research, developing biennial per capita income estimates, evaluating the quality of decennial census coverage, and for statistical research and development projects for the Census Bureau's current surveys (for example, evaluating the quality of income data reported in the Current Population Survey). The demographic area's use of these individual income tax returns (Form 1040), for research and evaluation associated with the decennial censuses, dates back to the 1960s. Samples of persons listed on individual tax returns have been matched to decennial census records to evaluate coverage. Although the coverage evaluation program for 1990 has not been thoroughly developed, it may very well include matches of IRS and Census Bureau records.

We have used the individual return information for the population estimates work since 1972. The Population Estimates Program uses concurrent year tax files to produce migration matrices representing the change in population for individual governmental units from one year to the next. Together with birth and death records, these estimates of population change are used each year to update population estimates for states, counties, and subcounty governments.

The biennial per capita income (PCI) estimates for states, counties, and local governments are developed simultaneously with the population estimates. Five tax return income items, including adjusted gross income, dividends, interest, wages and salaries, and gross rents and royalties are used in developing the PCI estimates. Estimates of income changes between two tax years are used to update the PCI estimates.

2. PROBLEMS ENCOUNTERED IN USING TAX DATA

Since we began using IRS data, there have been occasions when, because of misunderstanding, poor communication, or lack of adequate control procedures, the files received from the IRS were not what we expected. For instance, in the 1982 Economic Censuses, the Census Bureau requested that the IRS add several questions to the 1982 business tax forms. One of these questions never made it to the form. Also, requests for regulation changes to allow the Census Bureau access to 1982 tax information were delayed nearly three years from the original request due to inadequate coordination. Finally, industry coding was seriously deficient due to shortcuts taken at the regional processing centers to meet tough production quotas. These problems have not been the norm, however, and, were not devastating to the 1982 Economic Censuses program. Nevertheless, this was a costly lesson to learn. The IRS and the Census Bureau shared responsibility for these mistakes and are committed to making the 1987 Economic Censuses run more smoothly. This is being done by establishing quality and timing standards; holding regular review meetings; and developing a computerized monitoring system to track progress and issue regular status reports (Jonas, Hanczaryk 1987). The major overhaul of the tax system, effective in tax year 1987, makes this effort all the more challenging.

3. RESULTS OF OUR EVALUATION

Given the Census Bureau's reliance on tax information, the proximity to the Census Bureau's economic censuses and the decennial census, and the kinds of problems that can occur when using administrative records, the process of evaluating the effects of the 1986 Tax Reform Act becomes all the more urgent. In order to take charge of this situation and ensure a coordinated effort, the Census Bureau's Program and Policy Development Office began, in late 1986, to study the effects of the 1986 Tax Reform Act on Census Bureau programs. This study involved understanding the Act and its implications for Census Bureau programs; discussing these implications with the Statistics of Income (SOI) Division at the IRS and Census Bureau Divisions using these data; monitoring all tax form and rules changes; understanding the processing implications in terms of content and quality of the files we receive; and tracking processing operations to identify backlogs. This is being accomplished in cooperation with the SOI Division, which responds to our questions regarding tax reform implementation and provides us with draft copies of all forms and regular processing reports. As a backup, we review all **Federal Register** notices for IRS requests, made to the Office of Management and Budget, to clear new and revised 1987 tax forms and any rules and procedures that implement the new law. In addition, we have sought out published information on the 1986 Tax Reform Act for any relevance to Census Bureau programs and future needs. As a result of this effort and our ongoing quality assurance programs, we are increasingly confident that we are not going to repeat the mistakes of 1982 and that we will be able to take full advantage of the statistical potential of the 1986 Tax Reform Act while minimizing the negative effects on our programs.

The process of evaluating the effects of tax reform is ongoing. The 1987 tax forms are in various stages of completion and the processing system is being rewritten to handle the changes. We have identified anticipated benefits and losses to our programs and are evaluating the less obvious consequences as a result of changes in taxpayer recordkeeping practices and behavioral reponses.

3.1 New Information

In terms of benefits, the 1986 Tax Reform Act presents several potential opportunities to the Census Bureau. For example, Section 1524 requires that taxpayers must now obtain

and report social security numbers for child dependents age five and older. Any decennial census coverage evaluation activities involving direct matches of IRS and Census Bureau records could make use of this information.

Specifically, teenage children appear to be a group most likely to be reported on more than one return (their own and their parent's). The availability of social security numbers would assist the Census Bureau in unduplicating these cases prior to matching.

Another statistical benefit from tax reform is a requirement created by Section 1521 of the Act that gross proceeds of real estate transactions be reported to the IRS. One of the parties responsible for the closing is required to provide, on new form 1099-S, a description of the property (address or legal description), the gross proceeds from the transaction excluding the value of property or services received, and the date of closing.

As a part of each quinquennial Census of Governments, the Census Bureau has conducted a Taxable Property Values Survey (TPVS). Each such survey since 1957 has had two major components. The first is an assessment-sales price ratio study, the only source of actual property tax assessment levels that is nationwide in coverage. The second component is an estimation of the magnitude and composition of the real property tax base in terms of assessed value and numbers of parcels.

For 1987, tight budget have forced us to drop the ratio study and concentrate exclusively on tax base composition. In this setting, the new real estate sales price reporting requirements of IRS present some interesting possiblities. With a few modifications, such as a requirement that the assessed value and the existing use of the same property also be reported, Form 1099-S could well become a vehicle for collecting the data necessary to calculate assessment-sales price ratios. Conceivably this augmented form could not only help IRS in its tax collection efforts, but also accomplish a lot more. The Census Bureau could use the information for nationwide ratio studies on an annual rather than quinquennial basis. Similarly, the many states that conduct their own ratio studies could potentially use information from the 1099-S and eliminate the need for their own independent collection efforts. We are currently discussing with IRS the desirability and feasibility of modifying Form 1099-S so as to enhance its usefulness for administrative, as well as statistical purposes.

Finally, "taxpayers will (now) be required to report all tax-exempt interest received. As a result of this and other provisions, it will be possible, relying only on information reported on tax returns, to more closely approximate standard economic concepts of income" (Jabine 1987). As mentioned previously, the Census Bureau currently receives five income items from the IRS for use in evaluation studies and to produce the per capita income estimates. Expanding the IRS interest item to include nontaxable interest would provide better comparisons with interest income reported in the March Supplement to the Current Population Survey and the Survey of Income and Program Participation (SIPP) and would provide long-term benefits to the per capita income estimates.

3.2 Information We Will Lose

On the downside, there is some information, useful in our demographic statistics programs, that will no longer be collected as a result of tax reform. For instance, tax records for persons age 65 and older are excluded from our county-level population estimates procedure because tax file coverage for this group is not considered adequate — many persons over 65 do not file returns. Other sources, such as medicare files, provide better coverage. Since age is not required on tax returns, there is no direct way to identify this group so that we can exclude them from our procedure. However, before the 1986 Tax Reform Act, persons aged 65 and older were entitled to an additional personal exemption. Therefore, we were able to determine if the tax filer was over 65 by the presence of this exemption. Under tax reform, this exemption was eliminated and

replaced with an extra standard deduction, but only for filers 65+ who do not itemize. For itemizers, we cannot directly infer age. For this group, presence of social security income on the tax return could be used as a correlate variable for an age allocation procedure. This would only serve as a proxy, however, for determining age as under 65, since persons under 62 can receive social security income if they have a disability as can persons aged 62-64. Filing of Schedule R (credit for the elderly) could also be used in such a procedure.

In some preliminary research done by the Census Bureau using the 1984 SOI sample file, we found that we could capture 64 percent of those persons over 65 using the presence of an extra standard deduction for the elderly. By defining also the primary and secondary filer as 65+ if they had some social security income, we could correctly pick up an additional 29 percent, but would erroneously pick up an additional 17 percent. On the other hand, if we define only the primary filer as 65+ if there was some social security income, we correctly pick up an additional 21 percent while erroneously adding 8 percent. The presence of Schedule R had a negligible contribution (less than 1 percent) to the identification of those over 65. Research continues on this and other age correlates to approximate the age information lost because of tax reform (Sater 1987).

Perhaps a more dramatic loss of information will result from the large number of taxpayers who are no longer required to pay taxes because of the increase in the personal exemption allowance and the amount of standard deduction. It has been estimated that nearly 5 million poor people will be removed from the tax rolls (Pechman 1987). The loss of these low-income persons could affect the migration data developed as part of the population estimates program. We are aware that the exclusion of nonfilers has some affect on our migration data. Assuming that low-income persons do not move as much as higher-income persons, migration data from tax returns may not accurately reflect true migration for the entire population. For example, we may overstate the degree of outmigration for certain large cities. Significant increases in nonfilers resulting from tax reform would add to this negative differential.

Based on a study done at the Census Bureau using three different tax filer simulation algorithms, and an adjusted 1984 SOI income file, we estimate that, at most, 4 to 5 percent of the returns filed in Tax Year 1986 will not be filed in Tax Year 1987 solely because of changes in filing requirements (Valdisera 1987). Furthermore, these potential nonfilers will represent only 2 to 3 percent of all persons (filers and dependents) in the 1986 tax filer universe. On the other hand, we expect the nonfiler estimates for specific geographic areas and demographic groups to vary considerably from this national estimate.

These projections are based on our experience that many potential nonfilers will continue to file a return either by choice, out of habit, out of ignorance of the new filing requirements, or out of failure to notify their employers to stop withholdings. We are continuing this research in the interest of improving our estimates of potential nonfilers. As these estimates increase, we will be looking toward an alternative source of migration information for the poor.

3.3 Other Factors

In addition to the information added or lost, there are several factors that affect the quality and timeliness of the information being reported and processed. For instance:

- As with any major processing system revision, delays could result when handling these extensive tax form changes. Program rewrites may not be available when they are needed and may not run smoothly, if they are ready.

- Clerical coding and keying backlogs, resulting from new procedures, may prevent timely delivery of files to the Census Bureau.

- Not all information collected on tax forms is keyed for 100 percent of the returns. Knowledge of what information is only partially keyed is critical in decisions to use tax data.

- The extent of key verification and editing done to the data varies depending on its administrative uses. The quality of the data and its usefulness for statistical programs will, therefore, be affected.

- Taxpayers may generally file late causing large processing workloads that cannot be handled quickly with available resources.

We have some indication that many taxpayers will indeed file late 1987 returns because of late or improper filing of Form W4, Employee's Withholding Allowance Certificate. The W4s, filed previously, tend to under-withhold earnings after accounting for the adjustments to the tax rate tables made in January 1987. Taxpayers were asked to file new W4s to account for changes in allowable deductions. However, confusion about this requirement and the complexity of the form caused the IRS to issue a revised form, extend the deadline for submitting the W4, and waive penalties in some instances. Depending on the extent of late W4 submissions, many taxpayers could be faced with owing additional taxes. Many of those who would usually file in January or February may delay filing until April.

Since our Population Estimates Program and the economic censuses programs depend on timely receipt of tax information, we will need to closely monitor workloads to spot potential bottlenecks. We have requested that the IRS provide us with copies of their weekly receipts and processing reports. By comparing these with previous year reports, we should be able to detect problems and make the necessary allowances.

Other problems could occur with the quality of the information received or processed. For instance, confusion over tax law changes could cause reporting problems, at least initially. On the other hand, with new compliance checks added to handle the changes, reporting may improve. At least for business tax return data, we hope to identify any changes and potential problems through our quality assurance program, which compares actual reporting patterns to established standards.

There are other changes that may have varying effects on the data: 1) Partnerships and S (small)-corporations* are now required to report earnings and expenses on a calendar, rather than fiscal year basis, meaning there will be more uniformity in reporting for these types of business returns; and 2) Businesses may disincorporate or elect to convert to an S-corporation in order to take advantage of the lower individual rates (28 percent maximum for individuals vs. 34 percent maximum for corporations). In the latter case, businesses that switch to S-corporations may apply for new Employer Identification Numbers (EINs), although there is no requirement that they do so. This may give a false measure of business births and deaths and could lead to a lag in receiving lists of true births, which are needed for sampling. Even if businesses do not apply for new EINs, these

* S-corporation is a legal form of organization established for "small" businesses to allow them to take advantage of the tax breaks afforded unincorporated businesses, yet providing the benefits of incorporation. In order to elect as an S-corporation the business must have fewer than 35 shareholders and all shareholders must agree to the election.

shifts, if significant, will result in considerable increase in part year returns that would, in turn, complicate census processing.

The decision to switch to an S-corporation is complicated, however, because of: 1) the 33 percent marginal tax rate imposed on individuals with taxable income between $71,900 and $149,250; 2) limitations placed on S-corporations in borrowing from pension plans, 3) restrictions on S-corporations in issuing stock; and 4) the requirement of S-corporations to report on a calendar year basis. "S-corporation status still might be best for certain types of companies. But it's not going to be an easy call" (Quinn 1987).

In 1985 there were 735,000 S-corporations and, previously, they had been increasing at about 30,000 per year. For tax year 1986, we estimate that there will be 950,000 S-corporations while only 770,000 had been projected. Based on elections to become an S-corporation that were received as of July 1987, the IRS estimates that there will be approximately 1.1 million S-corporations in tax year 1987. We intend to monitor these closely and take account of any substantial increases in S-corporations as we process the economic censuses.

As previously noted, another complication involves the behavioral reactions of taxpayers (who happen to be also Census Bureau respondents) to the limitations/opportunities that arise from tax reform. For instance, deferral of income and acceleration of deductions and capital gains will affect year-to-year comparisons of these items as reported to the Census Bureau. In addition, treatment of tax shelters, real estate writeoffs, and IRA provisions will affect investment in these types of assets. Accounting for behavioral changes amounts to keeping informed so that we can anticipate how our data may be affected.

4. WHAT THE FUTURE HOLDS

Dealing with tax reform is a reflection of the Census Bureau's goal to give appropriate attention to data we do not collect that are, nonetheless, a vital component of our operations. It is all too easy to assume that the administrative agency will maintain the integrity and quality of their data relative to our needs. What we can lose sight of is the fact that the administrative agency's mission is different from that of the statistical user. No matter how important we think our uses are, we remain a relatively insignificant component of the administrative agency's operations. Consequently, we need to make our needs known, keep informed, and plan for contingencies. We can do this by taking a more active role in seeking out changes, but we also need a commitment from the administrative agency to provide us with relevent system information on a timely basis. This balance is what we are looking for in all of our uses of administrative records for our statistical programs.

Looking beyond the 1986 tax law, and the tax data in general, the Census Bureau is developing a means of allowing all statisticians to evaluate major U.S. Federal administrative records systems for applicability to their statistical programs. This system, known as the Administrative Records Information System, or ARIS, is based on some preliminary work begun by the Subcommittee on Statistical Uses of Administrative Records of the U.S. Federal Committee on Statistical Methodology. It is intended to collect detailed information needed by statistical users in deciding which, if any, major Federal administrative records systems satisfy these statistical needs and what access and quality limitations will have to be dealt with.

We began by identifying major Federal systems that we wanted to include in our study. Using the results of Project Link-Link* to identify systems that had previously been used in linkage studies, major systems identified in Statistical Policy Working Paper No. 6 (see ref.), and other systems found in **Federal Information Sources and Systems** (see ref.), we targeted approximately 55 administrative systems controlled by 20 government agencies. Questionnaires were sent to each agency requesting information on:

- physical structure of the file, including file size;
- the population included on the file (e.g. individuals, households, businesses);
- what information is on the file;
- how the information is obtained;
- the frequency of collection, updating and correction of information;
- ability to access the information;
- studies on the quality of the data;
- availability of documentation; and
- who to contact for further information.

There was a great deal of support for this undertaking and we are quite pleased with the response. Needless to say there were many questions regarding how the information would be used and we met and talked by phone with agency representatives to explain our purpose and coordinate the reporting within the agency. In some cases we found that the system we had identified was not truly an administrative file or was a minor subset of a larger system. Occasionally, the agency explained that the files and information were confidential and protected by the Privacy Act or legal statute and could not be accessed by anyone other than agency personnel. Generally, we were able to reassure the agencies that those who used the database would be well aware of the limitations on access but that it was important to include their system for completeness.

To date, we have received 47 completed questionnaires that have been entered into an interactive, relational database. (See Table 2 for current list of files.) This database has been compiled to run directly from a DOS operating system on an IBM-compatible personal computer. We will be able to provide this database on floppy disks or the user can dial into a Census Bureau electronic bulletin board and download the file directly. For bulletin board access, the user needs a modem, communications software, and PC/MS-DOS version 2.0 or higher.

One of the most important features of this system is the ability to keep the information up-to-date. We have designed the bulletin board to allow the provider agencies to correct and update easily information that we will then carry to the database. In addition, we will, annually, send out printed versions of the database information to the agencies for verification. In this way, we hope to keep ARIS as current and useful as possible.

---

* Project LINK-LINK is a gathering of information on administrative record linkage studies. Information on 26 studies was collected in early 1985 through the efforts of the Administrative Records Subcommittee of the Federal Committee on Statistical Methodology. Each study consists of information on the purpose for the linkage, the methodology used, the files that were linked, legal considerations involved in the matching, how the data were disseminated, and persons to contact for further information. The information is contained in an interactive database and is available on-line through a Census Bureau electronic bulletin board or on floppy disks.

With increasing concern for minimizing burden on respondents and reducing costs, we must look to administrative records as a possible alternative or supplement to survey or census data collection efforts. We think that ARIS will give us a means of evaluating that alternative. In addition, it will provide us with important information on changes to the files we currently use. Finally, it will assist in efforts to measure and improve coverage in our surveys and censuses, to improve estimation techniques, and to evaluate and supplement survey data.

Table 2

Administrative Records Systems Contained in ARIS

1. THE NATIONAL DEATH INDEX
   National Center for Health Statistics
2. THE INDIVIDUAL INCOME TAX EXTRACT FILE
   Internal Revenue Service
3. THE W2/W2P WAGE AND TAX STATEMENT EXTRACT FILE
   Internal Revenue Service
4. STATISTICS OF INCOME - INDIVIDUAL INCOME TAX RETURNS
   Internal Revenue Service
5. COMPENSATION AND PENSION MASTER RECORD FILE
   Veterans Administration
6. THE LIST SAMPLING FRAME
   National Agricultural Statistical Service
7. THE BUSINESS MASTER FILE
   Internal Revenue Service
8. THE EMPLOYMENT/PAYROLL EXTRACT FILE (FORMS 941/943)
   Internal Revenue Service
9. STATISTICS OF INCOME DIVISION PARTNERSHIP SAMPLE FILE
   Internal Revenue Service
10. THE CHARACTERISTICS OF FOOD STAMP HOUSEHOLDS FILE
    Food and Nutrition Service
11. THE RETURN PEACE CORPS FILE
    Peace Corps
12. THE INDIAN HEALTH SERVICE HEALTH CARE STATISTICS SYSTEMS
    Indian Health Service
13. THE MASTER ESTABLISHMENT LIST
    U.S. Small Business Administration
14. THE U.S. ESTABLISHMENT AND ENTERPRISE MICRODATA FILE
    U.S. Small Business Administration
15. THE VA CHAPTER 106 EDUCATION MASTER FILE
    Veterans Administration
16. THE MASTER PROVIDER OF SERVICES FILE
    Health Care Financing Administration
17. THE CENTRAL PERSONNEL DATA FILE
    Office of Personnel Management

18. THE RESIDENTIAL ENERGY CONSUMPTION SURVEY
Energy Information Administration

19. THE 1980 DECENNIAL CENSUS 100% FILES
Bureau of the Census

20. The 1980 DECENNIAL CENSUS SAMPLE FILES
Bureau of the Census

21. THE INDIVIDUAL MASTER FILE
Internal Revenue Service

22. THE PAYER MASTER FILE
Internal Revenue Service

23. THE SUPPLEMENTAL SECURITY RECORD
Social Security Administration

24. THE SUPPLEMENTAL SECURITY RECORD
Social Security Administration

25. THE UNEMPLOYMENT INSURANCE NAME AND ADDRESS FILE
Bureau of Labor Statistics

26. THE SMALL BUSINESS ADMINISTRATION LOAN ACCOUNTING FILE
Small Business Administration

27. THE INFORMATION RETURNS PROGRAM FILE
Internal Revenue Service

28. THE HEALTH INSURANCE MASTER ENTITLEMENT's FILE
Health Care Financing Administration

29. THE GUARANTEED AND INSURED LOAN SYSTEM
Veterans Administration

30. THE DEFICIENCY, DISASTER, AND DIVERSION PAYMENTS SYSTEM
Agricultural Stablilization and Conservation Service

31. OFFICE OF GENERAL SALES MANAGER SYSTEM
USDA: Agricultural Stabilization and Conservation Service

32. THE STATISICS OF INCOME CORPORATE SAMPLE FILE
Internal Revenue Service

33. THE IRS ESTATE TAX RETURNS FILE
Internal Revenue Service

34. RAILROAD EMPLOYER'S CREDITABLE COMPENSATION RECORDING SYSTEM
U.S. Railroad Retirement Board

35. RAILROAD RETIREMENT, DISABILITY, AND SURVIVOR BENEFIT PAYMENT SYSTEM
U.S. Railroad Retirement Board

36. THE GENERAL REFUGEE FILE
Health and Human Services

37. THE STANDARD STATISTICAL ESTABLISHMENT LIST*
Bureau of the Census

38. SUMMARY EARNINGS FILE
Social Security Administration

39. THE MASTER BENEFICIARY RECORD

Social Security Administration

40. THE NUMIDENT FILE
Social Security Administration

41. THE EMPLOYER REPORT RECORD
Social Security Administration

42. THE SINGLE-UNIT CODE FILE
Social Security Administration

43. THE MULTI-UNIT CODE FILE
Social Security Administration

44. THE CHAPTER 30 EDUCATION MASTER FILE
Veterans Administration

45. THE CHAPTER 31 TARGET MASTER RECORD
Veterans Administration

46. THE CHAPTER 32 EDUCATION MASTER FILE
Veterans Administration

47. THE CHAPTER 34 and 35 EDUCATION MASTER FILE
Veterans Administration

REFERENCES

Comptroller General, Federal Information Sources and Systems 1984, Washington, U.S. Government Printing Office, 1984.

Jabine, T.B. (1987). "Statistical Uses of Administrative Records in the United States: Some Recent Developments". Paper presented at the Annual Meeting of the Statistical Society of Canada in Quebec.

Jonas, J., and Hanczaryk, P. (1987). "Automatic Quality Assurance Processing of Administrative Records Files" presented at Symposium on Statistical Uses of Administrative Data.

Pechman, J.A. (1987). Tax Reform: Theory and Practice; *Economic Perspectives*, Vol. 1, No. 1; 11-28.

Quinn, J.B. (1987). "The Pluses and Minuses of Becoming an "S" Corporation;" Washington Post — Business Section.

Sater, D. (1987). Internal memoranda, U.S. Bureau of the Census.

Statistical Policy Working Paper No. 6; Report on Statistical Uses of Administrative Records; U.S. Department of Commerce, Office of Federal Statistical Policy and Standards; December 1980.

Valdisera, V. (1987). Internal memorandum, U.S. Bureau of the Census.

Wakefield, J.C. (1987). *The Tax Reform Act of 1986*; Survey of Current Business; 21.

SESSION VI: CONTRIBUTED PAPERS

Chairperson: M.P. Singh, Statistics Canada

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

A REVIEW OF THE USE OF ADMINISTRATIVE RECORDS IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

CHESTER BOWIE and DANIEL KASPRZYK[1]

ABSTRACT

The Survey of Income and Program Participation (SIPP) is a new Census Bureau survey providing information on the social, demographic, and economic characteristics of the nation's persons and families. Data collected in the SIPP include: income sources and amounts, program participation, labor force activity, and types and amounts of assets. From its inception, the goal of the SIPP was to create a data system which integrated both administrative and household survey data. Although this goal has not yet been achieved, several projects have begun which intend to use the survey data with administrative data. This paper reviews the SIPP projects that plan to use data from administrative sources — 1) augmentation of the survey data with earnings history data and social security benefits data; 2) a record check study to evaluate responses to questions on selected types of income; and 3) estimation research using administrative record data.

INTRODUCTION

The Survey of Income and Program Participation (SIPP) is intended to provide comprehensive information on the economic resources of the American people and on how public transfer and tax programs affect their financial circumstances. The data from the SIPP are expected to provide government policy makers with an information base for studying the efficiency of government tax and transfer programs, for estimating future program costs and coverage, and for assessing the effects of proposed policy changes.

The SIPP provides comprehensive information about annual and sub-annual income and participation in public and private transfer programs for the household population in the United States; it also devotes considerable attention to measuring various kinds of economic resources other than current cash income. The most important elements of this broader perspective are the SIPP data on assets, debts, and non-cash resources, such as means-tested housing benefits, publicly and privately provided health insurance, pension coverage, and other employee fringe benefits.

The SIPP arose in response to the recognition that the principal source of information on the distribution of household and personal income in the United States — the March

[1] Chester E. Bowie, Chief, Income Surveys Branch, Demographic Surveys Division, Bureau of the Census, Washington, D.C. 20233,
Daniel Kasprzyk, Chief, SIPP Research and Coordination Staff, Office of the Director, Bureau of the Census, Washington, D.C. 20233.

Income Supplement of the Current Population Survey (CPS) — had limitations which could only be rectified by making substantial changes in the survey instrument and procedures. One of the limitations of the CPS was the inability to provide linkages to administrative record data for statistical purposes. Recognizing this limitation of the CPS and the analytic usefulness of linking survey data to administrative records, the designers of the SIPP explicitly stated the ultimate SIPP data system should be a combination of data from administrative records and household surveys linked through the Social Security Number. The goals of the SIPP as described by Lininger (1980) state that administrative records will be used to:

1. increase sampling efficiency for certain subpopulations (e.g., Old Age, Survivors and Disability Insurance recipients or Supplemental Security Income recipients);
2. compare with survey data for validation studies of items common to both sources; and
3. supplement survey reported data with administrative record data for items difficult to obtain in a survey (e.g., earnings and program benefit histories).

These goals manifest themselves first in the SIPP development program and then in the SIPP.

This paper describes the SIPP's continuing commitment to the use of administrative records for statistical purposes: Section I reviews the work of the research and development program preceding the SIPP with regard to the use of administrative record data; Section II describes the design and content of the SIPP and its program to obtain accurate reporting of the Social Security Numbers (SSN) to facilitate linkages between survey reported data and administrative record sources; Section III describes five areas of applications where the SIPP project has initiated work involving the use of administrative records for statistical purposes; and Section IV provides a few additional examples of potential administrative data — SIPP linkage possibilities.

1. THE USE OF ADMINISTRATIVE RECORDS IN THE ISDP

The Income Survey Development Program (ISDP), authorized in 1975, was a program whose goal was to develop methods and a survey design to overcome underreporting and misclassification problems in the CPS (Yeas and Lininger, 1981). Furthermore, the ISDP also developed procedures and methodology for improving the collection of SSNs. The philosophy, attitudes, and plans of the ISDP strongly reflected the work of Scheuren and his colleagues (Scheuren et al, 1975) in the development of the 1973 Exact Match File (Kilss and Scheuren, 1978). A review of the work of the ISDP with regard to the use of administrative records can be found in Kasprzyk (1983) and Griffith and Kasprzyk (1980). A brief summary of this work will provide a context for the discussion of the SIPP experience and the plans and potential for the statistical uses of administrative records in the SIPP.

In the ISDP the collection and accurate reporting of the Social Security Number (SSN) from each person in sample was deemed essential to the program. By emphasizing the collection of the SSN and then developing a system to validate and correct reported SSN's, 95.5% of the total cases were identified as having a correct SSN (Kasprzyk, 1983). The system served as a prototype for the SIPP system which is described in the next section.

The ISDP consisted of four experimental field tests which were conducted to examine different concepts, procedures and questionnaires. One aspect of each of these field tests was the use of an administrative record frame for sampling purposes. Even though the principal thrust of this approach was to increase sampling efficiency for selected

subpopulations through the use of multiple frame estimators, the most important result was that these feasibility studies provided an opportunity for the survey planners to understand the administrative, methodological and operational difficulties in using administrative sources for sampling.

During the ISDP the following administrative record sources were used: 1) the Aid to Families with Dependent Children (AFDC) master file maintained by the Texas State Department of Welfare[2], the Supplemental Security Record (SSR)[3], the Master Beneficiary Record (MBR)[4], the Basic Educational Opportunity Grant (BEOG) applicant file[5], the Veterans Administration Pension and Compensation file[6], the Internal Revenue Service Individual Master File[7], and State record files for Unemployment Insurance and Workers Compensation.

The ISDP also effectively used administrative records to clarify misreporting and nonreporting of program benefits by comparing the survey reported data with the administrative data. Vaughan (1978) and Goudreau, Oberheu and Vaughan (1981, 1984) report on ISDP studies which led to redesigning questionnaires in order to reduce errors in classifying sources of income.

Finally, although the ISDP intended to create a data base augmented with administrative data which were difficult to obtain in a household survey it never did. A planned match to the Summary Earnings Record[8] was never implemented because of higher priority projects.

---

[2] The Aid to Families with Dependent Children (AFDC) master file is an administrative system maintained by each individual State and containing data on benefit amounts, payment history, demographic characteristics, and other information needed to administer the program.

[3] The Supplemental Security Record (SSR) is the national master administrative file for data on Supplemental Security Income (SSI) benefit amounts, payment history, and demographic data.

[4] The Master Beneficiary Record (MBR) is the national master administrative file for data on the Old Age, Survivors, and Disability Insurance Program (Title II); it contains current and historical program information on claimants for Title II benefits, past and present cash beneficiaries, disallowed claimants, and denied claims.

[5] The Basic Educational Opportunity Grant (BEOG) file is an administrative file maintained by the Department of Education. It contains data for all applicants of a given academic year, including ineligibles, eligibles who did not use their grant, and eligibles who used their grant. The BEOG program is now called the Pell Grant program.

[6] The Veteran's Administration Pension and Compensation File is a national master file containing records of benefits provided as disability compensation, dependency and indemnity compensation, disability pension, death pension, or burial allowance.

[7] The Internal Revenue Service Individual Master File (IMF) is a national file of selected income and tax information from all individual Income Tax Returns pertaining to wages, dividend and interest income, taxes paid, and exemptions.

[8] The Summary Earnings Record (SER) is a file containing the lifetime covered earnings (up to the maximum for each employer) and quarters of social security coverage of the individual. It is used to determine entitlement to benefits and calculation of benefit amounts. Individuals are identified in this file by their Social Security Number.

2. SIPP: DESIGN AND CONTENT

This section provides a very general summary of the SIPP survey design and content, followed by a description of the program established to collect and validate Social Security Numbers.

1. SIPP Design Features

The primary goals in designing the SIPP were to improve reporting of income and other program-related data and to do it in a way that would allow the analysis of changes over time at a microlevel. The design also had to accommodate the collection of a large quantity of information in a flexible manner that allowed some information to be collected more frequently than other information. These goals were met principally by using a survey design in which the same people are interviewed more than once. Persons (15 years of age or older) at households selected for a sample panel are interviewed about their income and other topics once every 4 months for approximately 2 1/2 years. Sample persons are interviewed at new addresses if they move, and any other persons that they move in with, or vice versa, are also interviewed. In this way, a highly detailed record is built up over time for each person and household in a sample panel. This design minimizes the need for sample persons to recall most of the information for longer than a few months and reduces the number of questions asked in one interview.

To further enhance the estimates of change, particularly year-to-year change, a new sample panel is introduced every year instead of at the conclusion of a panel. Consequently, two and sometimes three panels are in the field concurrently. The overlapping panel design allows cross-sectional estimates to be produced from a larger, combined sample that is about double in size when 2 panels overlap and triple with 3 overlapping panels.

The first SIPP panel, designated as the 1984 Panel but fielded in October 1983, started with approximately 20,000 interviewed households. The second panel, i.e., the 1985 Panel, began in February 1985 with around 14,000 interviewed households. Panels of about 12,300 interviewed households are expected to be fielded every February. The sample size changes in each wave of a panel due to losses through attrition and gains from following movers to new households.

The reference period for the primary survey items is the 4 months preceding the interview; for example, in February, the reference period is the preceding October through January. When the household is interviewed again in June, the reference period is February through May. To create manageable interviewing and processing work loads each month instead of one large work load every 4 months, the sample households within a given **panel** are divided into four subsamples of nearly equal size. These subsamples are called **rotation groups**, and one rotation group or one-fourth of the sample is interviewed each month. Thus, it takes 4 consecutive months to interview the entire sample. This 4-month period of interviewing is called a "**wave.**"

2. SIPP Survey Content

Each interview is planned to take about 30 minutes of a respondent's time and includes content that is divided into three main groups of questions. The substance of two of these groups should be essentially the same for each wave and for each panel. The third group of questions covers topics that will change in each wave of a panel. This allows for the inclusion of some new content in each panel, although many of the topics will be repeated across all the panels. Each rotation group in a wave is administered the same set of questions although the reference period is different as explained above.

The first group of questions are control card items. The control card is a separate document from the questionnaire and serves several important functions. The control card is used to list every person residing at an address and to record basic social and demographic characteristics (age, race, sex, and so forth) for each person at the time of the initial interview. Some information relating to the housing unit or household also is collected; e.g., number of units in the structure, tenure, and so forth. The card is reused at subsequent interviews to record changes in characteristics such as age, educational attainment, and marital status, and to record the dates when persons enter or leave the household. Finally, during each interview, information on each source of income received and the name of each job or business is transcribed to the card so that this information can be used in the updating process at the next interview.

The second major group of questions form the core portion of the questionnaire, which is divided into 5 sections. The core set of questions is asked at the first interview and then updated in each subsequent interview. The first section of the core collects the basic labor force participation data for the 4 reference months.

In addition, this first section of the core collects much of the information on the receipt of income from various sources during the 4 month reference period. This includes income from government sources such as Aid to Families with Dependent Children, Supplemental Security Income, General Assistance, and Workmen's Compensation. Respondents are also asked about both Social Security and other retirement income including Railroad Retirement, pension from company or union, and civil service retirements, as well as others. The receipt of miscellaneous sources of income such as alimony, child support, interest from savings, income for foster child care, and educational assistance is also identified. In addition, questions on major sources of noncash benefits such as food stamps, WIC (Women, Infants, and Children Nutrition Program), Medicaid, Medicare, and health insurance coverage are included in this section.

The second section of the SIPP core questionnaire collects information associated with wage and salary earnings. This section includes information on industry and occupation as well as hourly earnings for up to two jobs.

The third section of the core collects data on self-employment earnings and specific information about the kind of self-employment — whether it was incorporated, sole proprietorship, or partnership — and the profits and losses from the business. Again, space is provided for two self-employment jobs.

The fourth section is identified as the general amounts section. This section of the questionnaire collects monthly amounts received from the income sources identified in the first section. That is, the first section identifies the receipt of income during the 4 month reference period, while amounts of income received are collected in the fourth section of the questionnaire. Space is provided for amounts from up to six income sources.

The fifth and last section of the core questionnaire collects amounts of income earned from asset holdings. Asset sources include savings accounts, bonds, stocks, and rental property, as well as others. Information is collected for the 4 month reference period on both individual and joint recipiency.

The third major question grouping consists of the various supplements or topical modules that are included in waves following the initial interview. A wide variety of topics are covered under the aegis of the topical module concept. The breadth of these data ensure that SIPP will be a widely used and powerful data base serving multiple purposes. The administration of a module is possible in Waves 2 through 8 (or 9 in 1984) because less time is required to update the core information after the first interview. Depending on the time available and length of the modules, more than one may be administered in the same wave. The topical modules cover areas that do not require examination every 4 months and may use a different reference period than the core

questions. Some modules are assigned to only one wave of a panel, while other modules may be repeated in more than one wave. The modules provide a broader context for analysis by obtaining information on a variety of topics not covered in the core portion of the questionnaire. The module data may be analyzed independently or in conjunction with the control card items or core data. Frequently, a module is administered at the same time in concurrent panels so that the data may be combined to improve reliability.

3. The Collection and Validation of Social Security Numbers in the SIPP

The SIPP data system has always been thought of as a combination of data from administrative records and household surveys. This reduces respondent burden by using other data sources for difficult-to-obtain information. Interview responses can be supplemented by information from program files such as the earnings and benefit records of the Social Security Administration (SSA). This allows, for example, analysis of the long-term impact of various Social Security benefit formulas.

To make these linkages accurate, Social Security Numbers (SSN) are required for sample individuals. The SSN is obtained for each household member in SIPP and recorded on the control card. It is identified as a critical survey data item requiring completion to make the interviewers aware of its importance. These numbers are then verified and corrected to maximize the number of accurate linkages to other record systems.

The verification and correction process builds on the work of the development program (Kasprzyk, 1983). At the conclusion of each month's interviewing during the first wave of a SIPP panel, a special extract file is prepared by the Census Bureau for the SSA. This file contains a small number of key variables (SSN, name, date of birth, sex) for all original sample persons who report a SSN, including children, in a format appropriate for machine validation. Persons who report that they do not have a number or have a number but cannot supply it are handled separately in a clerical (manual) procedure.

Persons who refuse to provide a SSN are not included in the search process. The SSA identifies (by machine validation) incorrectly reported numbers then clerically resolves these cases along with cases not reporting a SSN. This work is completed by the fourth wave interview, at which time a field followup is conducted to obtain missing SSNs (provided they are not "refusals") and to reconcile inconsistencies in SSN or demographic data generated by the computer match or the clerical resolution.

Social Security Numbers of persons who enter the sample after Wave 1 (because they start living with original sample people) are validated at the start of the next panel. For example, information on new panel members (nonsample persons) from Waves 2 through 5 of the 1984 Panel was held and submitted for computer validation with Wave 1 of the 1985 Panel. Likewise, information on nonsample persons from Waves 6 through 8 of the 1984 Panel and Waves 2 through 4 of the 1985 Panel were held and submitted for computer validation with Wave 1 of the 1986 Panel.

The following summarizes the SSN validation results for the 1984 Panel Wave 1 sample:

53,588	Total Wave 1 sample persons
-1,674	Persons who refused to provide a SSN and were excluded from the validation process
51,914	Persons eligible for SSN validation
-42,128	Persons who reported a usable SSN and were eligible for computer validation
9,786	Persons who did not report a SSN and were eligible for the manual search (mostly children)

44,172	Validated SSNs (85% of eligible)
7,742	Unvalidated SSNs (mostly children who have no SSN)
51,914	Eligible for SSN validation

Based on these results, Sater (1986) has concluded that the SSN acquisition rate for persons who have a SSN is between 93 and 97 percent.

3. SIPP LINKAGE: WITH ADMINISTRATIVE DATA

This section briefly describes five areas of application where work has begun to use the survey data and administrative record data in some capacity: 1) SIPP/SSA data linkage project; 2) Employer Provided Benefits Study; 3) SIPP Record Check Study; 4) the use of Administrative Records in SIPP Estimation; and 5) merging economic data with SIPP demographic data.

1. SSA/SIPP Data Linkage Project

SSA's interest in a data set which merges administrative data with household survey data follows closely the intended uses of SIPP at its inception. A merged data set would enable the SSA to:

1. Estimate future program costs — The SSA is responsible for projecting program costs for all major SSA programs including: The Old Age, Survivor, and Disability Program, the Supplemental Security Income Program and Aid to Families with Dependent Children. In order to improve the accuracy of the projection methods, the SIPP panel data can be linked with a number of years of SSA data so that inflows and outflows can be analyzed in addition to point-in-time prevalence estimates of SSA program participation. The relationship between program participation and underlying individual characteristics can then be used to estimate future program costs and growth thus providing the SSA an early forecasting capability.
2. Assess the effects of program policy changes — An SSA-SIPP linkage will contain family, income and SSA benefit data. This combination of information will permit the SSA to estimate the programmatic costs of policy changes that depend on these factors and to assess the effects of policy changes on the economic well-being of program participants.
3. Describe non-programmatic characteristics of program participants — The SSA is frequently asked by Congress and others to provide information about program participants that is not routinely captured by administrative record systems. In the past, the SSA has used a series of widely spaced and usually one-time surveys to provide such information. Since the prospects for a new round of special purpose surveys are "not good", an ongoing SSA-SIPP data link would provide relatively up-to-date data on a routine basis.
4. Test social science theories as they relate to Social Security programs — The longitudinal component of the SIPP's research design and the wealth of data captured in core questions and topical modules provide data that will be sufficiently rich to test many social and economic theories of program participation, thus making a significant contribution to the basic research that must accompany any dynamic social program.

In essence, the project involves a maximum linkage with SIPP. For each SIPP panel, all waves of data, including core questions and topical modules will be linked to extracts of the basic SSA program records: The Master Beneficiary Record (MBR) which contains eligibiliity and benefit histories of the OASDI program, the Supplemental Security Record (SSR) which contains eligibility and benefit histories for the SSI program, and the Summary Earnings Record (SER) which contains a history of covered earnings for each worker. SSA records will be updated periodically so that each panel's files will contain additional years of the SSA's program data. We may also want to link to new disability administrative files that are now being developed at the SSA on a regular basis. All initial and subsequent linkages will be by mutual agreement between the SSA and the Bureau of the Census.

The merged data set will reside on the computer at the Census Bureau and will be used only for general statistical research. Only Census Bureau staff and SSA employees who are designated as Census Bureau Special Sworn Employees will have access to the file. The SSA may publish statistical data in a summary form that does not permit the identification of a household, family, or individual.

The primary tasks in the linkage project are:

1. Verification of, and searching for, Social Security Numbers (SSN) — This task is already a part of the SIPP project activities and was described earlier. In particular, the vast majority of SSN's for the 1984 SIPP panel have been processed by SSA staff.
2. Obtaining SSA administrative records — As mentioned above, this project involves matching the MBR, SSR and SER to the SIPP. Decisions will have to be made about the content of the data extracts from these files that would be included in the match.
3. Merging administrative records with SIPP survey data — The matching tasks are not one-time activities. Instead we anticipate a number of data processing operations for each SIPP panel.
4. Weighting, imputation and sampling error estimation — We will have to consider and develop schemes for weighting and imputation that take into account non-matched SIPP records. Both cross-sectional and longitudinal weights will be required. The SSA would also need the capability for estimating sampling errors.
5. Development of documentation for the matched files — Documentation for a matched file would include tape description and utilization information, the SIPP questionnaires and descriptions of the SSA administrative records, a sampling statement, imputation descriptions and any other information required for estimation or analysis.

2. Employer Provided Benefits Feasibility Study

Employer contributions to health insurance plans, retirement plans and life insurance plans have recently been the focus of national attention on the part of Congress, other policy makers, and researchers in areas such as health care, the elderly, and tax reform. SIPP collects information on whether a person is covered by health insurance and whether the employer makes contributions, but stops short of obtaining amounts for either the respondent's contribution or the employer's contribution. For life insurance, information is obtained on coverage, face value, and whether policies are provided through an employer. Amounts of employee payments and employer contributions are not obtained.

This study involves obtaining a signed release from the respondent at the interview and contacting the respondent's employer and asking the employer to fill out a short questionnaire to obtain data on both the employer's and employee's contributions to health insurance plans, pension plans, and life insurance plans. Information provided by the employer would supplement the SIPP data.

A half sample of one rotation group's households was selected for the study. The test was done in August 1987, (rotation group 4) for households in Wave 8 of the 1985 Panel. This was the last interview for these households.

The test included only employed persons, 18 years old and older, for whom a Wave 8 interview questionnaire was completed. Of the 1,352 persons eligible for the test, 569 persons (42 percent) signed the authorization form, 446 persons (33 percent) refused to sign, and 337 proxy or telephone respondents (25 percent) did not return the authorization form that was left/mailed to them. We did not conduct a followup of the refused or non-return cases.

Of the 569 questionnaires that were mailed to an employer, 548 (96 percent) were completed and returned. A more detailed evaluation of the data collected in this study will be undertaken next year, together with an assessment of the future prospects for a study of this type on the complete sample.

3. SIPP Record Check Study

Another area of research with respect to administrative record systems is the development of validation studies of items common to both the survey and administrative records. The purpose of the study is to investigate response quality issues in SIPP through a case-by-case comparison of SIPP data and administrative record information. The ultimate goal is the improved understanding of the quality of the SIPP data and, ultimately, the development of quantitative estimates of response and nonresponse errors for the purposes of adjusting survey data or modifying survey procedures to obtain better quality survey data.

An overview and progress report of the study can be found in Moore and Marquis (1987). Simply put the study intends to address the following questions:

1. The quality of the respondent reports of receipt of program benefits for a variety of state and Federally administered transfer programs;
2. The quality of benefit dollar amount reporting for these programs;
3. Demographic correlates of report quality;
4. Extent of misclassification errors;
5. The (nonexperimental) effects of self-proxy respondent status on report quality; and
6. Between wave recipiency turnover effects (The "seam" problem (Burkhead and Coder, 1985; Moore and Kasprzyk, 1984)).

The questions will be addressed by using administrative record information for recipients of each of nine government transfer programs in four states--Florida, New York, Pennsylvania, and Wisconsin. Four state-administered programs (Aid to Families with Dependent Children, food stamps, unemployment compensation, and worker's compensation) and five Federally-administered programs (Civil-Service Retirement, Pell Grants, Old Age Survivors and Disability Insurance (OASDI), Supplemental Security Income, and Veterans' Pensions and Compensation) will be studied. The project has obtained a great deal of information on acquiring administrative record systems, learning about each system's idiosyncrasies, and generalized matching procedures at the Census Bureau. Some very preliminary results are now available in Moore and Marquis (1987).

4. Use of Administrative Records in SIPP Estimation

Information on the effect of sample reductions on the variance of estimates and on our ability to measure changes in differences in a number of statistics have created serious concerns. These concerns have caused us to increase our exploration of ways to reduce the variance. One approach is through the use of administrative records for post-stratification. Currently, cross-section estimation procedures for SIPP make use of a second-stage adjustment to increase the precision of estimates by ratio adjusting collection month and reference month estimates to population estimates. However, the Census Bureau has access to some Internal Revenue Service and Social Security Administration files which can be used to produce detailed age, race, and sex distributions by adjusted gross income. The issue, which we have just begun to explore, is how these administrative data can be used for post-stratification to improve estimates of mean and median personal and household income as well as the estimates of the deciles of the personal and household income distribution. Furthermore, a basic question which will be considered is how much reduction in the variances of these estimates can be achieved through such a procedure. These issues will be investigated next year.

The first phase of this research (Huggins, 1987) will estimate the reductions in variances of SIPP estimates by using the IRS data as auxilliary variables in the estimation procedures. The procedure being studied has been advocated by Herriot (1985) and Scheuren (1983). In the SIPP study the estimation method will involve a ratio adjustment of SIPP estimates at the second stage of estimation in cells defined by age + race + sex + "income" where "income" is adjusted gross income as reported to the Internal Revenue Service.

Controls are prepared from a 1% sample of 1984 IRS file matched with age, race, and sex characteristics from the Summary Earnings Record; adjusted gross income from the 100% IRS file is matched to a file of SIPP data. The SIPP cases are then reweighted by controlling to the 1984 IRS controls; that is, a factor f_j, which is the ratio of IRS control in $cell_j$ to the SIPP estimate of persons matched to IRS data with 1984 IRS income in $cell_j$, is applied to persons who fall in $cell_j$ based on the IRS data. Estimates and variances of selected SIPP characteristics will be obtained using the newly created weights and with the weights which do not use this procedure.

5. Merging Economic Data with SIPP Demographic Data

During the first two years of the SIPP program a good deal of background research was completed on the potential for augmenting SIPP data with micro-level establishment and enterprise data from the economic census and other data files maintained by the Bureau of the Census (Haber, Ryscavage, Sater, Valdisera, 1984). Haber (1985) has described the analytic potential of matching economic data to the demographic data for individuals in the SIPP. Haber suggests that new insights are possible in the following areas: the relationship between capital and wage rates, the study of labor mobility between low and high-wage employees, studying implications of the transition from goods-producing to a service economy, and analyzing the effects of unions on the labor market. A pilot project was initiated to investigate methodologies for merging individuals in the SIPP (who report their employer's name) to the employer data in the economic census, testing the methodology to identify problem areas and solutions, and conduct the match for a pilot sample. Sater (1985) describes the project, and problems encountered. Unfortunately, due to costs, higher priorities, and staffing limitations, this project was never completed.

3. POTENTIAL LINKAGE TO OTHER ADMINISTRATIVE DATA SETS

The SIPP is a relatively new continuous survey, collecting a comprehensive socio-economic portrait of the U.S. household population. As mentioned above, the SIPP also gives substantial attention to the correct reporting of Social Security Numbers. These two elements together provide the principal reasons for the power of the data set. In the future, the good match variable (SSN) which the SIPP provides could be used in matching the survey data to the Health Care Financing Administration's Health Insurance Master File (Medicare) to study the relationship between hospital use, health status, employment and income. Similarly, the SSN will allow linkage of deceased respondents to the National Death Index. In the latter case, numerous SIPP panels would be necessary to have sufficient sample for analysis. Nevertheless, the potential for such linkages exist. In fact any linkage with an administrative record system which uses the SSN as the primary identifier is possible. The principal difficulties, however, are the costs for such projects and the difficulty of sharing matched administrative-survey data with all researchers.

REFERENCES

Burkhead, D., and Coder, J. (1985). Gross Changes in Income Recipiency from the Survey of Income and Program Participation. *Proceedings of the Social Statistics Section, American Statistical Association*, 351-356.

Goudreau, K., Oberheu, H., and Vaughan, D. (1981). An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 377-382.

Goudreau, K., Oberheu, H., and Vaughan, D. (1984). An Assessment of the Quality of Survey Reports of Income from the Aid to Families with Dependent Children (AFDC) Program. *Journal of Business and Economic Statistics*, 179- 186.

Griffith, J., and Kasprzyk, D. (1980). The Use of Administrative Records in the Survey of Income and Program Participation. Case study in *Report on Statistical Uses of Administrative Records: Statistical Policy Working Paper 6*. U.S. Government Printing Office, Washington, D.C. 20402.

Haber, S., Ryscavage, P., Sater, D., and Valdisera, V. (1984). Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study. *Proceedings of the Social Statistics Section, American Statistical Association*, 529-533.

Haber, S. (1985). Applications of a Matched File Linking the Bureau of the Census Survey of Income and Program Participation and Economic Data. *SIPP Working Paper Series No. 8502*, U.S. Bureau of the Census, Washington, D.C.

Herriot, R. (1983). The Use of Administrative Records in Social and Demographic Statistics. Paper presented at the Meeting of the International Statistics Institute, Madrid, Spain.

Huggins, V. (1987). Research Plans. Memorandum for the Record, April 13,1987, Statistical Methods Division, U.S. Bureau of the Census.

Kasprzyk, D. (1983). Social Security Number Reporting, the Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program in *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program (ISDP)*, M. David (editor), 123-141. New York: Social Science Research Council.

Kilss, B., and Scheuren, F. (1978). The 1973 CPS-IRS-SSA Exact Match Study. Social Security Bulletin, Vol. 41, No. 10, 14-22.

Lininger, C. (1980). The Goals and Objectives of the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 480-485.

Moore, J., and Kasprzyk, D. (1984). Month-to-Month Recipiency Turnover in the ISDP. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 726-731.

Moore, J., and Marquis, K. (1987). Using Administrative Record Data to Evaluate the Quality of Survey Estimates. Paper presented at the International Symposium on the Statistical Uses of Administration Records, November 23-25, 1987, Ottawa, Canada.

Sater, D.K. (1985). Enhancing Data from the Survey of Income and Program Participation with Data from Economic Census and Surveys. *SIPP Working Paper Series No. 8505*, U.S. Bureau of the Census, Washington, D.C.

Sater, D.K. (1986). SSN Response Rates and Results of SSN Validation/Improvement Operation. Memorandum for Roger Herriot, March 11, 1986, Population Division, U.S. Bureau of the Census.

Scheuren, F., Herriot, R., Vogel, L, Vaughan, D, Kilss, B., Tyler, B., Cobleigh, C., and Alvey, W. (1975). Report No. 4: Exact Match Research using the March 1973 Current Population Survey--Initial States. *Studies from Interagency Data Linkages.* U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics, Department of Health, Education and Welfare, publication No. SSA 76-11750.

Scheuren, F. (1983). Design and Estimation for Large Federal Surveys Using Administrative Records. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 377-381.

Vaughan, D. (1978). Errors in Reporting Supplemental Security Income Recipiency in a Pilot Household Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 288-293.

Yeas, M., and Lininger, C. (1981). The Income Survey Development Program: Design Features and Initial Findings. *Social Security Bulletin*, Vol. 44, No. 11, 13-19.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

TIME SERIES MODELLING FOR SMALL AREA ESTIMATION

G.H. CHOUDHRY and L.A. HUNTER[1]

ABSTRACT

Most government statistical agencies, such as Statistics Canada, carry out continuous surveys, and hence, time series data for small areas is available. Generally, the data for small areas will not be of acceptable reliability, so a time series modelling approach can be considered to obtain smoothed estimates for small areas with improved reliability. The auxiliary information which enters into this time series model is based on the local data which includes survey estimates, and data from recent census and administrative records. In our application we have used the Canadian Labour Force Survey (LFS) as an example, and have used time series modelling with auto-regressive errors, to obtain smoothed estimates of unemployment for census divisions.

1. INTRODUCTION

In recent years, the demand for more timely and reliable small area data has been steadily increasing due to their use in formulating policies and programs, in allocation of government funds, and in regional planning. Statistics Canada responded to the user needs by undertaking a program of small area data development. The issues arising in the development and provision of small area data are discussed by Brackstone (1986).

The full range of small area data demands cannot be met using direct estimation from sample surveys because of cost and response burden implications due to the large sample sizes required to produce reliable small area estimates. It is therefore important to develop methods of estimation that combine data from existing sample surveys, recent census, and administrative records using statistical models. Such methods include synthetic estimators (e.g. Gonzalez and Hoza; 1978, and Ghangurde and Singh; 1978), and sample size dependent estimators due to Drew, Singh and Choudhry (1982), and Hidiroglou and Särndal (1985). But none of these methods exploit the information due to the correlation in the time series data to obtain more reliable small area estimates. In this paper, we use the pooled cross-sectional time series data in a regression model with an auto-regressive error structure. Taking the Canadian Labour Force Survey (LFS) as an example, we have used the model to obtain smoothed estimates of unemployment for census divisions. The auxiliary information which enters into this model is based on the local data which includes survey estimates, data from the most recent census and administrative counts from the unemployment insurance system. As expected, the time series modelling technique resulted in estimates of improved reliability over those that do not exploit the time dimension.

[1] G.H. Choudhry and L.A. Hunter, Social Survey Methods Division, Statistics Canada.

The paper is structured as follows. The basic model is presented in section 2, and the procedure for estimating the model parameters is described in section 3. In section 4, the model is evaluated for estimating unemployment for census divisions. In section 5, various alternatives for updating the model over time are considered and finally a few concluding remarks are made in section 6.

2. BASIC MODEL

The field of survey statistics usually deals with cross-sectional data describing each of many different units (e.g. small areas) at a single point in time, whereas econometrics usually uses time series data describing a single entity. Since most large scale surveys are continuous surveys, a cross-sectional time series data base is also available for small areas. This data can be pooled to obtain smoothed estimates for small areas. Pooling consists of combining cross-sectional and time series data on the variables for estimating the coefficients of the regression model. Although separate regression models can be estimated for the individual areas, pooling over areas results in more efficient estimation of the model if a common underlying model for small areas, possibly with different intercepts, can be identified. This type of model for pooled cross-sectional and time series data has been considered by Dielman (1983). In our application, we have used the cross-sectional time series data to fit a regression model in order to improve the reliability of the unemployment estimates for census divisons. Cronkite (1984) has used separate regression models for obtaining state employment and unemployment estimates, but for local area employment and unemployment estimates developed a model by combining time series data cross-sectionally to form a pool of observations for estimation purposes. Binder and Dick (1987) have also followed a similar modelling approach for small area estimation in the context of the Canadian Travel Survey.

At this point we introduce the following notation. For small area 'a' at time t, let ${}_ay_t$ be the observation on the dependent variable and ${}_aX_{kt}$ (k=0, 1, ..., K) be the observations on the explanatory variables. There are K+1 of these variables and k=0 corresponds to the constant term. We also introduce the dummy variables ${}_aZ_{a'}$, which are equal to one if a is equal to a' and zero otherwise. Then the basic time series model is of the form

$$ {}_ay_t = \sum_{k=0}^{K} \beta_k({}_aX_{kt}) + \sum_{a'=2}^{A} \gamma_{a'} ({}_aZ_{a'}) + {}_au_t, $$

$$ \begin{aligned} a &= 1, 2, \ldots A \\ t &= 1, 2, \ldots T \end{aligned} \qquad (2.1) $$

The β's and γ's are the regression coefficient and ${}_au_t$ are the stochastic errors. The assumed error structure for the model (2.1) is

$$ {}_au_t = \rho\ {}_au_{t-1} + {}_a\varepsilon_t $$

where ρ is the auto-correlation coefficient and is assumed to be the same for all small areas and across time. Furthermore we assume that the ${}_a\varepsilon$'s are independently and identically distributed with mean zero and variance ${}_a\sigma^2$ for a=1, 2, ... A.

The ${}_ay_t$ is the natural logarithm of the LFS estimate of the proportion of unemployed to population. The transformation to logarithms makes the errors more symmetric and

homoskedastic and also reduces the impact of extreme values. The ${}_aX_{kt}$'s are the (K+1) predictors including the constant term.

The K explanatory variables depend on area and/or time, e.g. the administrative counts from the UI system and the LFS estimates depend on both area and time, the census variables depend on area only, and month and year dummy are time-dependent. The dummy variables ${}_aZ_{a'}$ are introduced to account for different intercepts for small areas. The error term ${}_au_t$ includes the error due to sampling and also reflects the correlated structure of the estimates over time within small areas. All variables except the dummy variables are expressed as proportions or rates, or functions thereof.

3. ESTIMATION

First a stepwise regression was used for variable selection with the exception of the intercept term, LFS participation rate, and the logarithm of proportion of population receiving UI benefits, which were always included in the model. The LFS participation rate was a very good predictor of seasonal fluctuations in unemployment and hence was included in the model. To obtain efficient estimates of the model parameters, the following transformation as proposed by Cochran and Orcutt (1949) was used for the model (2.1).

$$({}_ay_t - \rho\ {}_ay_{t-1}) = \sum_{k=0}^{K} \beta_k\ ({}_aX_{kt} - \rho\ {}_aX_{k,t-1})$$

$$+ \sum_{a'=2}^{A} \gamma^{\star}_{a'}\ ({}_aZ_{a'}) + {}_a\varepsilon_t \tag{3.1}$$

where $\gamma^{\star}_{a'} = (1 - \rho)\ \gamma_{a'}$.

It was possible to define the above transformation for $t \geq 1$ as the data for t=0 were available. Alternatively, had t=0 not been available, the transformation for t=1 would have been obtained by multiplying equation (2.1) by $(1 - \rho^2)^{\frac{1}{2}}$ on both sides. However, this was not done, because it would have complicated the computational procedures. The transformed model (3.1) was specified in the form

$${}_ay_t = \rho\ {}_ay_{t-1} + \sum_{K=0}^{K} \beta_k({}_aX_t - \rho\ {}_aX_{t-1})$$

$$+ \sum_{a'=2}^{A} \gamma^{\star}_{a'}\ ({}_aZ_{a'}) + {}_a\varepsilon_t \tag{3.2}$$

and the least-squares estimates of the model parameters were obtained using the modified Gauss-Newton method (Hartley; 1961). These estimates will be referred to as unweighted least-squares estimates. From the estimated residuals, we obtained estimates ${}_a\hat{\sigma}^2$ of ${}_a\sigma^2$ for a = 1, 2, ..., A. We then used weighted least-squares for obtaining the model parameters where the weights were equal to ${}_a\hat{\sigma}^{-2}$. Again Hartley's modified Gauss-Newton method was used for estimating the model parameters. These estimates will be referred to as weighted least-squares estimates. From these estimates, the residuals for the transformed model (3.2) and the variances ${}_a\sigma^2$ were estimated. The residuals were tested for zero mean using t-tests for each of the small areas separately. If the

t-statistic was significant at the 5% signficance level for any of the small areas, then the dummy variable for that area was included in the model and the weighted least-squares estimates were obtained by repeating the procedure. This estimation procedure is virtually identical with the Aitken's generalized least-squares procedure.

Following Goldberger (1962), the best linear unbiased predictor (BLUP) of ${}_a y_t$ was obtained as

$$ {}_a\tilde{y}_t = \sum_{k=0}^{K} \hat{\beta}_k \, {}_a X_{kt} + \sum_{a'=2}^{A} \hat{\gamma}_{a'} \, ({}_a Z_{a'}) $$

$$ + \hat{\rho} \left\{ {}_a y_{t-1} - \sum_{k=0}^{K} \hat{\beta}_k \, {}_a X_{k,\, t-1} - \sum_{a'=2}^{A} \hat{\gamma}_{a'} ({}_a Z_{a'}) \right\}, \qquad (3.3) $$

where $\hat{\beta}$'s, $\hat{\gamma}$'s, and $\hat{\rho}$ are the least-squares estimates of the corresponding model parameters. It should be noted that

$$ \hat{\gamma}_{a'} = \frac{\hat{\gamma}^{*}_{a'}}{1 - \hat{\rho}} . $$

The estimate given by (3.3) is the sum of the systematic and the time series component. The additional gain in the efficiency is realized due to the time series component over the traditional estimator given by the systematic component (Goldberger; 1962). It should also be pointed out that the time series modelling approach is not used for the purpose of prediction outside the range of the model, rather it is used to obtain smoothed estimates with improved reliability. The model-based estimates of the proportions unemployed are obtained as

$$ {}_a\tilde{p}_t = \mathrm{Exp}\ ({}_a\tilde{y}_t) $$

with an associated variance given by

$$ v({}_a\tilde{p}_t) = \mathrm{Exp}\ (2\ {}_a\tilde{y}_t)\ v({}_a\tilde{y}_t) $$

for a = 1, 2, ...A and t = 1, 2, ...T where $v({}_a\tilde{y}_t)$ is the estimated variance of ${}_a\tilde{y}_t$. The coefficient of variation (CV) of ${}_a\tilde{p}_t$ can be obtained as

$$ CV({}_a\tilde{p}_t) = [\ v({}_a\tilde{y}_t)\ ]^{\frac{1}{2}}. $$

Thus the CV of the model-based estimate of unemployed is obtained simply by taking the square root of the estimated variance of the predicted value of the dependent variable in the model which is the logarithm of the proportion unemployed. Details of computation of $v({}_a\tilde{y}_t)$ are given in the Appendix.

The smoothed estimates of the level of unemployment are obtained by multiplying the estimated proportion unemployed by post-censal population estimates for small areas obtained from external sources (Verma et. al.; 1983).

The variance of the smoothed estimates of unemployment averaged over months (e.g. quarterly, annual, etc.) can also be obtained by making use of the variance-covariance matrix of the estimates ${}_a\tilde{y}_t$, t=1, 2, ...T, also given in the Appendix.

4. EVALUATION OF THE MODEL

Due to differences in the labour market conditions, separate models were estimated for the 5 Census regions: (i) Maritimes, (ii) Quebec, (iii) Ontario, (iv) Prairies, and (v) British Columbia. The sixth region, which comprises Yukon and Northwest Territories, could not be included because the LFS is not conducted in the territories. For each of the five regions, regression models were obtained using 36 months (January 83 - December 85) of time series data. The small areas in our application were the census divisions or groupings of census divisions. The number of small areas in each of the regions for which the data was pooled for the purpose of estimation, is given in Table 4.1.

Table 4.1

R^2 Values for Unweighted and Weighted Regressions and the Number of Small Areas by Census Region

Census Region	Number of Small Areas	R^2 Value for Unweighted Regression	R^2 Value for Weighted Regression
1	43	76.1	97.2
2	42	69.7	96.6
3	37	70.2	97.6
4	44	69.3	98.3
5	21	64.4	97.7

The R^2 value is a measure of the goodness-of-fit of the model and these were obtained for the unweighted and weighted regressions for each of the regions. These are also reported in Table 4.1. We observe that there is significant improvement in the goodness-of-fit due to weighted regression over the unweighted regression. This improvement is due to large variation in the estimated variances of the small areas. Also, the high values of R^2 for the weighted regression indicates that the proposed model gives an adequate fit for the data.

From the weighted regression, we obtained the model-based estimates of the unemployment and evaluated the consistency of these estimates against the post-stratified domain estimates. The post-stratified domain estimates were obtained by ratio-adjusting the value of estimated unemployed from the LFS in the small area by the ratio of the post-censal population estimate for the small area to the LFS estimated population in the small area. The post-stratified domain estimator is unbiased except for the ratio estimation bias which is negligible for large sample sizes. From the monthly post-stratified domain estimates, quarterly, annual, and three-year average estimates were produced for each of the small areas and these were compared with the corresponding model-based estimates. The absolute relative deviation (ARD) between the two monthly estimates for area 'a' was defined as

$$_a(\text{ARD}) = \frac{100}{T} \sum_t \left| \frac{_a\hat{u}_t - {_a\tilde{u}_t}}{_a\tilde{u}_t} \right|,$$

where $_a\hat{u}_t$ and $_a\tilde{u}_t$ are respectively the post-stratified domain and model-based estimates for small area a and time t and T is the number of time periods. The average ARD over all the areas within a region was then defined as

$$\overline{\text{ARD}} = \frac{1}{A} \sum_a {_a(\text{ARD})}.$$

The average ARD's were also defined for the quarterly, annual, and three-year average estimates in a similar fashion. The average ARD's for the monthly, quarterly, annual and three-year average estimates are given in Table 4.2 for each of the regions. We notice that for each of the regions, the average ARD's decrease monotonically as the averaging period increases which indicates that there is no systematic bias in the model-based estimates of the series. Moreover, as mentioned in the previous section, the t-test for the residuals from the final model was not significant at the 5% level for any of the small areas. Thus there is no evidence that the model-based estimates are subject to any systematic bias.

Table 4.2

Average ARD's between the monthly, quarterly, annual and three-year average post-stratified domain and model-based estimates by region.

Census Regions	Average ARD for			
	Monthly Estimates	Quarterly Estimates	Annual Estimates	Three-Year Average Estimates
1	15.3	8.6	4.5	2.5
2	18.4	10.7	6.5	3.4
3	15.9	9.3	4.9	2.7
4	21.6	12.2	7.3	4.2
5	13.9	8.5	4.8	1.5

There is substantial improvement in the reliability of the model-based estimates due to: (i) pooling cross-sectional data over small areas for estimation, and (ii) exploiting the information due to the correlation in the time series. Drew, Singh and Choudhry (1982) proposed a sample dependent estimator which is a linear combination of the post-stratified domain and synthetic estimators, where the weight on the synthetic component depends on the outcome of the sample. If the sample in the small area is "sufficient", then the weight on the synthetic component is zero. As the sample in the small area decreases, the reliance on the synthetic component increases. The weight on the synthetic component becomes equal to one when there is no sample in the small area. We have compared the average CV's of the three-year average model-based estimates with those of the sample dependent estimates. This comparison is shown in Table 4.3. We observe that the use of the time series model resulted in a reduction by factors of 0.4-0.5 in the estimated CV's over the sample dependent estimators for the 5 regions. The estimated autocorrelation is also given in the same table for each of the regions with the corresponding standard deviation (S.D.) in the parentheses. The high autocorrelations account partly for large gains for the model-based estimates using time series approach.

Table 4.3

Average CV's for the three-year average model-based and Sample dependent estimates and the estimated auto-correlation ($\hat{\rho}$) by region

Region	$\hat{\rho}$(S.D.)	Estimated CV. for Model-based	Estimated CV. for Sample dependent	Ratio
1	0.60(0.02)	3.0	7.2	0.4
2	0.56(0.02)	3.4	7.3	0.5
3	0.53(0.02)	2.8	6.6	0.4
4	0.49(0.02)	3.6	8.2	0.4
5	0.53(0.03)	2.4	5.8	0.4

5. MODEL UPDATING

If this methodology is to be used on a regular basis to obtain time series of smoothed estimates, it is necessary to determine how well the model performs over time and whether or not some form of updating the model should be used. If updating is required, it must also be determined how often the updating procedure should be applied.

To answer these questions, the "best" models for three different time periods (81-83, 82-84, 83-85) were fitted to Region 5 (B.C.) data. The variables selected for the 81-83 model and their estimated coefficients were then used on the 82-84 and 83-85 data to determine how reliable the estimates are if the same model is used without any updating. One way of updating the model without entirely re-fitting it is to use the same variables selected at an earlier time, but re-estimating the values of the regression coefficients. To test this, the variables selected for the 81-83 model were used with the 82-84 and 83-85 data, and new regression coefficients were estimated for each. The average ARD was calculated for the estimates under each of the models and these ARD's were compared to determine the effect of no updating vs. updating the coefficients vs. re-fitting the model by reselecting the variables. The results are shown in Table 5.1.

Table 5.1

Effects of Model Updating on B.C. Data

Model	ARD Three-Year Average
Fitted Model for 81-83	2.9
Updating Model for 82-84	
(1) Same variables and co-efficient values as 81-83	5.1
(2) Same variables as 81-83; Re-estimated coefficients	3.2
(3) New Variable Selection	2.7
Updating Model for 83-85	
(1) Same variables and co-efficient values at 81-83	8.2
(2) Same variables as 81-83; Re-estimated coefficients	2.7
(3) New Variable Selection	1.5

If no updating is done, the model deteriorates quite rapidly. After one year the ARD is nearly double and after two years it is almost triple its value for the 81-83 data. If, however, the same variables are used in the model, but the regression coefficients are updated each year, there is a large gain in the reliability of the model. The ARD's for the two later years remain at about the same level as the ARD for the 81-83 data. Comparing the results for the updated models with those of the re-fitted (or "best") models shows that even two years later the ARD for the updated model is only slightly higher than that of the re-fitted model. These results indicate that while updating the model is necessary if it is to be used over a long period of time, it is not necessary to entirely re-fit the regression model as new data becomes available.

Another area of concern is the effect of using out-of-date Census data in the model. The census data on labour force status is collected only in the decennial censuses and is not generally available until about two years after collection. This means that the Census data used in the model may be as much as twelve years out-of-date. To investigate the effect this has on the estimates, we fitted the "best" model for B.C. for the 81-83 period using 1971 Census data and then using 1981 Census data, and compared the ARD's from each.

The three-year average ARD was 3.3% using 1971 Census data and 2.9% using 1981 Census data. This seems to indicate that the out-datedness of the Census data has very little effect on the reliability of the model. The Census-based variables used in the model are the proportions of the labour force population in each industry type, and these proportions have remained fairly stable over time.

Although the effects of updating have only been tested on B.C., the results are expected to be similar for the other regions.

6. CONCLUSIONS

By pooling over small areas, the time series modelling produces gains in efficiency over the sample dependent estimator because it exploits the correlation structure present in the time series. Moreover, there is no evidence of systematic bias in the model-based estimates.

Tests with one Census region have shown that the model need not be completely re-fitted as new data becomes available. Instead, the variables chosen for the model at one time period can be used for some time, with updates made to the values of the coefficients only. Tests in other regions should be made to verify these results, but it is expected that the findings would be similar.

The variances have been estimated under the assumption that variance remains constant over time for each of the small areas (i.e. ${}_a\sigma_t^2 = {}_a\sigma^2$ for $a = 1, 2, \ldots, A$ and $t = 1, 2, \ldots, T$). One area of future research could be to investigate this assumption using jackknife techniques (Wu; 1986).

APPENDIX

Variance-Covariance Matrix of the Model-Based Estimates

The time series model (2.1) can be written in the matrix form as:

$$\underset{\sim}{Y}^{(1,T)} = \underset{\sim}{X}^{(1,T)} \underset{\sim}{B} + \underset{\sim}{U}^{(1,T)} \tag{A1}$$

where $\underset{\sim}{Y}^{(1,T)}$ is the vector of n (=A·T) observations on the dependent variable for the T time periods t = 1, 2, ..., T and for the A small areas a = 1, 2, ..., A and $\underset{\sim}{X}^{(1,T)}$ is the corresponding matrix of observations on the predictor variables including the intercept term and the dummy variables. $\underset{\sim}{B}$ is the vector of unknown parameters of the model except for the autocorrelation parameter ρ. $\underset{\sim}{U}^{(1,T)}$ is the vector of stochastic errors which depend upon the autocorrelation ρ.

The autocorrelation is removed by linear transformation using the observations at time $t > 0$ for each of the small area. The linear transformation for each of the small area is given by the $T \times (T+1)$ transformation matrix $\underset{\sim}{R}_1$, defined as $R_1(t,t) = -\rho$, $R_1(t,t+1) = 1$ for t = 1, 2, ..., T, and zero otherwise.

Now define $\underset{\sim}{R} = \underset{\sim}{I}_A \otimes \underset{\sim}{R}_1$, where $\underset{\sim}{I}_A$ is the identity matrix of size A (the number of small areas) and $\otimes$ is the Kronecker product. The transformed model (4.3) can then be written as

$$\underset{\sim}{R}\ \underset{\sim}{Y}^{(0,T)} = \underset{\sim}{R}\ \underset{\sim}{X}^{(0,T)} \underset{\sim}{B} + \underset{\sim}{E} \tag{A2}$$

where $\underset{\sim}{E} = \underset{\sim}{R}\underset{\sim}{U}^{(0,T)}$, with $\underset{\sim}{Y}^{(0,T)}$, $\underset{\sim}{X}^{(0,T)}$, and $\underset{\sim}{U}^{(0,T)}$ defined similar to $\underset{\sim}{Y}^{(1,T)}$, $\underset{\sim}{X}^{(1,T)}$, and $\underset{\sim}{U}^{(1,T)}$. The variance-covariance matrix of the errors of the transformed model (A2) is given by $\underset{\sim}{\Sigma} = \underset{\sim}{\sigma} \otimes \underset{\sim}{I}_T$ where $\underset{\sim}{\sigma}$ is the diagnonal matrix of dimension A with a-the diagonal element equal to $_a\sigma^2$, and $\underset{\sim}{I}_T$ is the identity matrix of size T.

Now define $\underset{\sim}{X}^* = \underset{\sim}{\Sigma}^{-\frac{1}{2}}\ \underset{\sim}{R}\ \underset{\sim}{X}^{(0,T)}$ and $\underset{\sim}{Y}^* = \underset{\sim}{\Sigma}^{-\frac{1}{2}}\ \underset{\sim}{R}\ \underset{\sim}{Y}^{(0,T)}$. Then we have that $\hat{\underset{\sim}{B}} = (X^{*\prime}\ X^*)^{-1}\ (X^{*\prime}\ Y^*)$. Note that the transformation matrix $\underset{\sim}{R}$ is evaluated at $\rho = \hat{\rho}$.

Let $\tilde{\underset{\sim}{Y}}(1,T)$ be the vector of model-based estimates given in (3.3), then $\tilde{\underset{\sim}{Y}}(1,T)$ can be written as

$$\tilde{\underset{\sim}{Y}}(1,T) = \underset{\sim}{R}\ \underset{\sim}{X}^{(0,T)}\ \hat{\underset{\sim}{B}} + \underset{\sim}{Y}^{(1,T)} - \underset{\sim}{R}\ \underset{\sim}{Y}^{(0,T)}$$

$$= \underset{\sim}{R}\ \underset{\sim}{X}^{(0,T)}\ \hat{\underset{\sim}{B}} + \rho\ \underset{\sim}{Y}^{(0,T-1)}$$

where $\underset{\sim}{Y}^{(0,T-1)}$ is defined similar to $\underset{\sim}{Y}^{(1,T)}$.

Now

$\text{var}\ (\hat{\underset{\sim}{B}}) = (\underset{\sim}{X}^{\star\prime}\ \underset{\sim}{X}^{\star})^{-1}$, and

$$\text{var}(\underset{\sim}{Y}^{(0,T-1)}) = \text{var}(\underset{\sim}{U}^{(0,T-1)}) = \frac{1}{1-\rho^2}\ \text{var}(\underset{\sim}{E}) = \frac{1}{1-\rho^2}\ \underset{\sim}{\Sigma}\ .$$

and the variance-covariance matrix of $\underset{\sim}{Y}^{(1,T)}$ can be written in a straight forward manner as

$$\text{var}(\tilde{\underset{\sim}{Y}}^{(1,T)}) = \underset{\sim}{\Sigma}^{1/2}\ \underset{\sim}{X}^{\star}\ (\underset{\sim}{X}^{\star\prime}\ \underset{\sim}{X}^{\star})^{-1}\ \underset{\sim}{X}^{\star\prime}\ \underset{\sim}{\Sigma}^{1/2} + \frac{\rho^2}{1-\rho^2}\ \underset{\sim}{\Sigma}\ .$$

Considering the a-th small area, let ${}_a\tilde{\underset{\sim}{Y}}(1,T)$ be the model-based estimate for the a-th small area, then

$$\text{var}({}_a\tilde{\underset{\sim}{Y}}^{(1,T)}) = \{{}_aX^{\star}\ (\underset{\sim}{X}^{\star\prime}\ \underset{\sim}{X}^{\star})^{-1}\ {}_aX^{\star\prime} + \frac{\rho^2}{1-\rho^2}\ \underset{\sim}{I}_T\}\ {}_a\sigma^2\ ,$$

where ${}_aX^{\star}$ is the sub-matrix of $\underset{\sim}{X}^{\star}$ corresponding to the a-th small area and $\underset{\sim}{I}_T$ is the identity matrix of size T. The variance of ${}_a\tilde{\underset{\sim}{Y}}(1,T)$ is estimated by substituting the estimates of ρ and ${}_a\sigma^2$.

ACKNOWLEDGEMENT

The authors wish to acknowledge the many useful comments from colleagues at Statistics Canada.

REFERENCES

Binder, D.A., and Dick, J.P. (1987), "Estimation and Modelling in Repeated Surveys," Internal Working Paper, Social Survey Methods Division, Statistics Canada.

Brackstone, G.J. (1986), "Small Area Data: Policy Issues and Technical Challenges," *Small Area Statistics, An International Symposium* (R. Platek, J.N.K. Rao, C.E. Sarndal, and M.P. Singh, eds.), John Wiley and Sons, 3-20.

Cochran, W.G., and Orcutt, G.H. (1949), "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44, 32-61.

Cronkite, F.R. (1984), "A Proposed New System for Developing State and Area Employment and Unemployment Estimates: An Overview," Internal Technical Report, Bureau of the Labour Statistics, Washington, D.C.

Dileman, T.E. (1983), "Pooled Cross-Sectional and Time Series Data: A Survey of Current Statistical Methodology," *The American Statistician*, 37-111-122.

Drew, J.D., Singh, M.P., and Choudhry, G.H. (1982), "Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey," *Survey Methodology*, 8, 17-47.

Ghangurde, P.D., and Singh, M.P. (1978), "Evaluation of Efficiency of Synthetic Estimates," *Proceedings of the American Statistical Association, Social Statistics Section*, 52-61.

Goldberger, A.S. (1962), "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57, 369-375.

Gonzalez, M.E., and Hoza, C. (1978), "Small Area Estimation with Application to Unemployment and Housing Estimates," *Journal of the American Statistical Association*, 73, 7-15.

Hartley, H.O. (1961) "The Modified Gauss-Newton Method for the Fitting of Non-Linear Regression Functions by Least Squares," *Technometrics*, 3, 269-280

Hidiroglou, M.A., and Särndal, C.E. (1985), "An Empirical Study of Some Regression Estimators for Small Domains," *Survey Methodology*, 11, 65-77.

Verma, R.B.P., Basavarajappa, K.G., and Bender, R. (1983), "Regression Estimates of Population for Sub-provincial areas in Canada," *Survey Methodology Journal*, 9(2), 219-240.

Wu, C.F.J. (1986), "Jackknife, Bootstrap and other Resampling Methods in Regression Analysis," *Annals of Statistics*, 14, 1261-1294.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

TURNING THE TABLES: IMPUTING FOR ITEM NONRESPONSE WHEN DONORS ARE SCARCE

JOHN L. CZAJKA[1]

ABSTRACT

"Hot deck" or other substitution methods of imputation for item response are applied typically in settings where the ratio of potential donors to records requiring imputation is quite large. However, the U.S. Internal Revenue Service has employed imputation in conjunction with double sampling to reduce supplemental data collection needed to edit corporation tax returns for statistical purposes. In this application, records requiring imputation outnumber potential donors by as much as nine to one. This paper discusses the problems posed by the scarcity of donors and reviews several modifications introduced into the methodology this year to improve the estimates of corporation statistics for minor industries. The methodology can be applied to other settings where administrative data must be edited extensively prior to statistical use.

1. INTRODUCTION

Converting data from administrative records to a form useful for statistical purposes can be a sizable undertaking with a substantial price tag. This observation certainly holds true for the work of the Statistics of Income (SOI) Division of the U.S. Internal Revenue Service (IRS), which produces analytical data files by extracting information from the tax returns filed each year by individuals and corporations. The records sampled by the SOI Division undergo extensive editing, which may involve collecting and processing additional information available only from the returns themselves.

To reduce one component of the data file production costs in its corporation income statistics program, IRS has employed a combination of double sampling and imputation in place of full editing. These procedures have been applied to a set of "catch-all" items for which a taxpayer reporting a nonzero amount must append a supplementary schedule detailing the sources of the amount. SOI editors reviewing the schedules often find that portions of what is reported as, let us say, Other Income, can be reclassified under more specific items (for example, Gross Receipts or Other Dividends), which provide more information to analysts. They edit the data fields, moving out some of what was originally reported as Other Income and adding it to one or more of these other fields. Rather than review all of the schedules associated with the sampled returns, IRS has done the following for a subset of schedules:

1. review the schedules for all "large" returns

[1] John L. Czajka, Senior Researcher, Mathematica Policy Research, Inc., 600 Maryland Ave., S.W., Suite 550, Washington, D.C. 20024, U.S.A.

2. review the schedules for selected smaller returns--those whose likelihood of being changed by an edit are high relative to other returns from the same major industry

3. review the schedules for a random subsample of the remaining returns (20% of those in financial industries and 10% of those in nonfinancial industries)

4. impute edit outcomes to the remaining returns, using the random subsample as donors

These procedures were first implemented for seven schedules in tax years 1981 and 1982. The methodology employed in these two applications has been described by Hinkins (1983, 1984), Czajka (1986, 1987), and Hinkins and Scheuren (1986). Following a two year hiatus, a modified imputation scheme has been employed with three schedules for tax year 1985 (Czajka, 1987).

The general approach is applicable to other administrative record data programs. In these days of belt tightening in government budgets, many agencies that develop administrative data files for analytical purposes could benefit from the application of similar techniques to their own data programs. This paper addresses a technical problem inherent in the use of imputation to compensate for data missing as the result of double sampling: namely, the small number of donors relative to the number of records being imputed.

2. THE USE OF DOUBLE SAMPLING AND IMPUTATION TO REDUCE EDITING

As presented by Cochran (1977), double sampling involves drawing a preliminary sample in order to obtain information on characteristics which are required to stratify the main sample. In the IRS application, the preliminary sample is also the primary sample, which is subsampled to obtain additional information from particular types of returns. The additional information is used to edit fields recorded for all returns in the primary sample. Double sampling reduces the data collection required to support the editing of these fields. Imputation of the missing information to returns not selected in the subsample makes it possible to produce complete micro records for all returns in the sample. However, this departure from the customary use of double sampling generates conditions which are unusual for applications of imputation as well.

2.1 Double Sampling to Obtain Edit Information

The role of double sampling in the corporation SOI program may be illustrated by showing how the review of the Other Income schedule may affect the final amount recorded for Other Income in the analytical file. The derivation of the final value of Other Income for an individual sample firm may be described as follows:

$$Y_i = B_i - C_i$$

where Y_i represents the final amount, B_i represents the beginning amount, and C_i represents the change. Ignoring weighting, the estimator of Other Income for an aggregate of sample firms is thus:

$$\Sigma Y_i = \Sigma B_i - \Sigma C_i$$

The aggregate estimate consists of the total beginning amount less the total amount removed. In the absence of editing, therefore, ΣC_i is the aggregate bias.

As a result of double sampling, we observe C_i for only one-fifth of the financial firms that were subsampled and only one-tenth of the nonfinancial firms. To estimate Other Income (or any of the other items affected by reviewing the Other Income schedule) for a population of firms that was subsampled, we could inflate each observed change by five if the return belonged to a financial firm and ten if the return belonged to a nonfinancial firm. Reweighting is a common method of estimation when double sampling has been employed. However, compensating for item nonresponse by reweighting is not the preferred approach when constructing a micro data file that may be used for purposes other than producing aggregate tabulations. Moreover, in this instance aggregate estimates are published for more than a thousand subpopulations. For many of the subpopulations reweighting would yield estimates of final amounts that, while unbiased, possess extremely high variance. The alternative approach of imputing the unobserved C_i produces complete micro records that greatly increase the utility of the data file. Furthermore, imputation offers greater flexibility to address the potentially high variance of estimates for detailed subpopulations.

Results of the imputations are summarized in Table 1 for Other Income, Other Deductions and Cost of Goods. Over the three items the use of double sampling and imputation substituted for the editing of 57,136 schedules. Changes were imputed to 25.9 percent of the items.

Table 1

Summary of 1985 Imputations to Other Income Other Deductions, and Cost of Goods

Item	Total Records Imputed	Records with Original Amount Changed by Imputation	Records with Originial Amount Not Changed by Imputation	Percentage of Records with Amount Changed by Imputation
Other Income	18,657	4,149	14,508	22.2%
Other Deductions	23,719	7,662	16,057	32.3%
Cost of Goods	14,760	2,982	11,778	20.2%
Total	57,136	14,793	42,343	25.9%

NOTE: All figures are unweighted values.

2.2 Imputing for Non-edits: The Scarce Donor Problem

Imputation methods based on the substitution of values from complete records to incomplete records have strong appeal because of the degree to which they can be automated and because they provide a means to generate internally consistent imputations of multiple items for individual records. Both the need to automate the process and the need to impute several items per record are characteristic of the SOI application, and they explain the reliance upon a modified "hot deck" approach in the 1981 and 1982 corporation imputations (Hinkins, 1984).

Typically, substitution methods are employed in contexts where the records with missing data on any given item or related group of items are small in number, relative to the records with complete data, and where the sample size is very large (numbering in the tens of thousands or more). The large pool of potential donors makes it possible to achieve close matches between donor and impute on characteristics related to the missing items. When the source of the missing data is double sampling, however, the relative numbers of records with complete versus incomplete data are reversed. In the IRS application being discussed here, the records requiring imputation outnumber the prospective donors by either four to one or nine to one--hence the allusion to "turning the tables" in the title of this paper. Despite the size of the corporation sample, the prospective donors for any of the three items number only two to three thousand.

The scarcity of donors creates several problems for the application of imputation methods based upon substitution. In addition to the variance introduced by double sampling, three problems are particularly noteworthy:

1. the ability to closely match each record to a similar donor is limited by the small number of donors
2. the usage of donors is potentially very unequal, further increasing the impact of imputation upon the variance of estimates
3. the risk of imputing implausible results is substantially increased

We discuss each problem in turn.

Substitution methods require large numbers of potential donors to enable the donor-impute matching to incorporate several covariates of the missing items. With few donors, it is possible to match records with respect to only a small number of characteristics broken down to very few categories. To place the problem in perspective, in the IRS example the degree of similarity between each donor-impute pair is limited by the fact that each of the two to three thousand donors must be matched to between four and nine records, on average. The mean square error at the micro level, therefore, will include significant bias as well as variance. The error introduced by imputation may remain significant up to high levels of aggregation.

In typical applications of substitution methods, where the number of potential donors greatly exceeds the number of records to be imputed, most donors are used either once or not at all; rarely is a donor used more than once. With the odds reversed however, there exists the potential for wide variation in the frequency of use of individual donors. Large disparities in donor use may increase the variance of estimation resulting from imputation. The imputation procedures employed by IRS in 1981 and 1982 included a feature intended to minimize the variation in donor use within adjustment cells. Donors were drawn in succession from the analog of a once-shuffled deck. Within an adjustment cell, therefore, any donor was used at most one more time than any other donor. To have much impact on the variation in donor use, however, this approach requires that the adjustment cells contain several donors each. When the number of donors is very small, this imposes a sharp restriciton on the number of adjustment cells that can be supported.

A problem exacerbated by the low ratio of donors to imputes is the possibility of imputing implausible values or outliers. The IRS experience provides relevant examples. In the context of the IRS application, implausible imputations include changes that an editor would be unlikely to make. Editors are unlikely to remove very small proportions from small original amounts or to remove large proportions from very large amounts, but there are no hard and fast rules that can be carried over to the imputation procedure or included in the consistency tests to which imputed records are subjected. In the 1981 and 1982 imputations, IRS relied initially on its adjustment cell design to ensure plausibile imputations. However, there were instances in 1982 where donors with small fractions

removed from large Other Income amounts were paired with records having small original Other Income amounts. Since proportionate rather than actual dollar changes are imputed, the net result was the imputation of exceedingly small changes--in several cases less than one dollar. The low ratio of donors to imputes contributed to this outcome by limiting the degree of resemblance that could be achieved between each record to be imputed and the donor it shared with several other records.

2.3 The 1985 Corporation Imputations: An Overview

In revising the imputation procedures for their use with the 1985 corporation statistics program, we introduced several modifications to address problems raised by the scarcity of donors. These new elements are summarized here.

The key element in the revised approach is the separation of the imputation of a change versus no change (or nonzero versus zero change) from the imputation of the conditional magnitude of the change. This separation of the two components of the imputed edit is both possible and effective because a substantial proportion of the edits to the donors result in no changes to the original amounts. (Recall that two groups of returns were not subjected to subsampling: large returns and returns for which the expectation was relatively high that editing would produce a change.) Separating the two components of the imputation permitted us to deal with the scarcity of donors in two ways: (1) impute the change/no change outcome from a probability matrix, and (2) impute the conditional magnitude of change within more appropriately defined adjustment cells.

From the standpoint of dealing with scarce donors, the advantage of imputing change/no change from a probability matrix is that we eliminate the need to rely on the explicit matching of records to individual donors. Consequently, we can smooth the change probabilities to reduce the variability due to sampling. This allows us to relax the constraint on minimum acceptable cell size and thus include more covariates or increase the number of categories in one or more dimensions of the matrix, thereby reducing the imputation bias.

Imputation of the magnitudes of nonzero changes still involves the matching of records to individual donors. However, this step utilizes only donors with changes, allowing us to define adjustment cells that meet minimum size criteria. In this way we may control the sampling variance of the imputed changes. Moreover, to the extent that the covariates of the **magnitude** of nonzero change differ from the covariates of the **probability** of nonzero change, we can reduce the imputation bias by specifying a different set of covariates than we use to define the probability matrix.

To reduce the likelihood of imputing implausible magnitudes, we matched donors and imputes within adjustment cells on the logged original amount of the item (i.e., Other Income, Other Deductions, or Cost of Goods). This direct solution to a problem observed in the earlier imputations had an unintended side effect--namely, it produced a substantial disparity in the frequency of use of donors within the same adjustment cell. This outcome is discussed below. The modifications summarized here are detailed in separate sections below and illustrated for Other Income.

3. IMPUTATION OF ZERO VERSUS NONZERO CHANGE

The binary choice between zero and nonzero change to Other Income was imputed from a probability matrix with three dimensions, representing classifications of returns by industry, size and the ratio of Other Income to Total Income. The observed probabilities calculated from the Other Income donors were smoothed to reduce the impact of sampling variation, as many of the cells were very small. Outcomes were imputed at random

within cells of the matrix, subject to the corresponding smoothed probabilities. The matrix and the smoothing algorithm are described below.

3.1 Dimensions of the Probability Matrix

In defining the dimensions of the probability matrix, we expanded the industry detail and refined the size classification relative to what had been used in 1981 and 1982. We also added a new covariate. The industry classification consisted of 23 groupings of minor industries, disaggregated from seven major industries. (See Appendix.) In its earlier imputations IRS recognized only ten industry classes. The increased detail is made possible by the decision to impute the change/no change outcome from smoothed probabilities.

Three size classes were defined for each of the 23 industry groupings, based upon the group-specific distribution of returns by assets and net income. In 1981 and 1982 IRS applied a uniform size classification to all returns. Substantial differences in the size distribution across industries argued for this revision.

Based upon the ratio of Other Income to Total Income, five additional classes were defined within each industry grouping. This ratio of income fields is the basis for IRS's selection of returns for editing rather than subsampling of the Other Income schedule, but it was not used previously in the imputation process. We included this "selection ratio" as a dimension of the probability matrix on the evidence of its strong covariation with the probability that Other Income was changed on a donor record in 1982.

Table 2 reports the mean probability of change to Other Income observed among donors classified by each of the three dimensions. The mean change probabilities by size class and selection ratio are reported separately for financial and nonfinancial firms. There is substantial variation along each of the three dimensions. With respect to the size class and selection ratio categories, however, it should be noted that the variation in the change probability among financial firms is small in comparison with what we observe among nonfinancial firms.

3.2 Smoothing of the Probability Matrix

The cell values of the 23x3x5 probability matrix were smoothed to reduce the variability due to sampling. The smoothing algorithm used the observed marginal shares of donor records with and without changes, by class within each of the three dimensions, to construct predicted numbers of records with and without changes **by cell.** From these predicted counts we then calculated predicted probabilities of change as ratios of predicted records with changes to the predicted records with and without changes. The smoothed probability for each cell was then calculated as a weighted sum of the observed probability and the predicted probability, with the weights being a function of the sampled number of donors in the cell.

Formally, the smoothing algorithm may be described as follows. Let N_{1ijk} represent the number of records in cell (i,j,k) with changes, and let N_{0ijk} represent the number without changes. The marginal share of records with changes in size class i=1 is:

$$SIZ_{11} = \sum_{jk}\sum N_{11jk} / \sum_{ijk}\sum\sum N_{1ijk}$$

Table 2

Probability of Change to Other Income Among Donor Records Classified by Industry, Size and Selection Ratio

Minor Industry			Size of Return			Selection Ratio		
Class	Prob. of Change	Number of Donors	Class	Prob. of Change	Number of Donors	Class	Prob. of Change	Number of Donors
			Nonfinancial Firms					
1	.211	90	1	.087	391	1	.047	492
2	.194	72	2	.163	787	2	.121	381
3	.259	216	3	.195	780	3	.190	410
4	.239	109				4	.207	372
5	.421	152				5	.297	303
6	.158	57						
7	.068	59						
8	.176	119						
9	.145	76						
10	.103	194						
11	.083	108						
12	.077	117						
13	.133	75						
14	.088	137						
15	.095	21						
16	.080	200						
17	.052	77						
18	.101	79						
			Financial Firms					
19	.958	289	1	.565	147	1	.552	181
20	.724	123	2	.638	268	2	.549	133
21	.260	50	3	.601	296	3	.596	151
22	.161	93				4	.714	119
23	.244	156				5	.661	127

Similarly, the marginal share of records **without changes** is:

$$SIZ_{01} = \sum_{jk}\sum N_{01jk} / \sum_{ijk}\sum\sum N_{0ijk}$$

By a parallel procedure we calculate marginal shares of records with changes by the other size classes, by selection ratio (SEL_{1j}), and by industry class (IND_{1k}). Likewise we calculate the shares of records without changes by size, selection ratio and industry classes.

Letting N_1 equal the total number of records with changes, we calculate the predicted number of records with changes in cell (i,j,k) as:

$$PRED_{1ijk} = N_1 * SIZ_{1i} * SEL_{1j} * IND_{1k}$$

and the predicted number without changes as:

$$PRED_{0ijk} = N_0 * SIZ_{0i} * SEL_{0j} * IND_{0k}$$

(Note that the sum of the $PRED_{1ijk}$ over i, j, and k is N_1 and, likewise, the sum of the $PRED_{0ijk}$ is N_0.) We then calculate the predicted probability of change in each cell as:

$$PROB_{ijk} = PRED_{1ijk} / (PRED_{1ijk} + PRED_{0ijk})$$

These predicted probabilities are used in combination with the observed probabilities to generate the smoothed probabilities.

If the cell sample size equals or exceeds a parameter value specified in the imputation program (we used size 50 in the 1985 application), then no smoothed value is substituted for the observed probability of change. If the cell sample size is below the parameter value, we calculate a weighted sum of the observed and predicted probabilities, where the weights are a function of the square root of the cell sample size. More specifically, the weight assigned to the observed probability is the square root of the ratio of the sample size to the parameter value. The weight assigned to the predicted probability is the difference between this fractional ratio and unity. Thus if a cell sample size were 25, the smoothed probability would be computed as 70.7 percent of the observed probability plus 29.3 percent of the predicted probability. If the cell size were only 10, the smoothed probability would be 44.7 percent of the observed probability plus 55.3 percent of the predicted probability. The reduction in the weight assigned to the observed probability as the cell sample size declines is inversely proportional to the increase in the standard error of the observed probability.

Table 3 presents the observed, predicted and smoothed probabilities calculated for Other Income donors in the major industry Wholesale Trade. Raw frequencies of total records and records with changes are reported as well. The sparse cell counts are typical of most of the major industries.

The benefits of smoothing are apparent. The very small numbers of donors in most cells impart substantial sampling variation to the observed probabilities of change. Smoothing the observed change probabilities reduces sharply the noise component in the between cell variation yet still leaves strong covariation between the change probabilities and the three dimensions of the matrix.

Clearly other smoothing algorithms could be applied--perhaps with better success. The algorithm demonstrated here is analogous but not identical to fitting a log-linear model consisting of the two-way interactions between the outcome change/no change and each of the three covariates. Smoothing on the basis of fitted log-linear cell frequencies is an obvious alternative that would merit consideration. The probable need to rely on adjustments to eliminate marginal sums of zero in some of the industries was the principal deterrent to further exploration of this approach during the development of the 1985 imputation methodology.

Table 3

Smoothing of Change Probabilities for Other Income:
Wholesale Trade Industries

Size Class	Selection Ratio Class 1	2	3	4	5	Selection Ratio Class 1	2	3	4	5
Minor Industry Group 1										
	Records with Changes/Total					**Observed Probabilities**				
1	0/4	0/4	0/2	0/4	0/0	.000	.000.	.000	.000	.000
2	1/11	2/9	0/4	3/9	3/4	.091	.222	.000	.333	.750
3	1/11	1/10	4/8	1/6	3/4	.091	.100	.500	.167	.750
	Predicted Probabilities					**Smoothed Probabilities**				
1	.029	.082	.097	.167	.294	.021	.059	.078	.120	.294
2	.070	.183	.213	.336	.512	.080	.200	.153	.335	.580
3	.082	.210	.243	.375	.554	.086	.161	.346	.303	.610
Minor Industry Group 2										
	Records with Changes/Total					**Observed Probabilities**				
1	0/8	0/3	0/4	1/2	0/1	.000	.000	.000	.500	.000
2	0/7	2/6	0/4	2/6	2/5	.000	.333	.000	.333	.400
3	1/2	2/8	1/7	1/6	2/3	.500	.250	.143	.167	.667
	Predicted Probabilities					**Smoothed Probabilities**				
1	.026	.074	.088	.154	.274	.016	.056	.063	.223	.235
2	.064	.168	.196	.314	.487	.040	.226	.141	.320	.459
3	.075	.193	.224	.351	.529	.160	.216	.194	.287	.562
Minor Industry Group 3										
	Records with Changes/Total					**Observed Probabilities**				
1	1/10	2/11	2/11	3/12	0/1	.100	.182	.182	.250	.000
2	2/25	4/16	7/25	8/14	3/11	.080	.250	.280	.571	.273
5	1/16	3/16	5/20	9/21	6/7	.062	.188	.250	.429	.857
	Predicted Probabilities					**Smoothed Probabilities**				
1	.038	.104	.123	.208	.353	.066	.141	.151	.229	.303
2	.090	.227	.262	.398	.579	.083	.240	.275	.490	.435
3	.105	.258	.295	.439	.619	.081	.218	.267	.432	.708

4. DEFINITION OF ADJUSTMENT CELLS FOR IMPUTING NONZERO MAGNITUDES

Research based on the 1982 corporation data did not identify strong covariates of the conditional magnitudes of change--in striking contrast to what we found for the probabilities of change. To impute the magnitudes, therefore, we dimensioned the adjustment cells similarly to the probability matrices, except that we excluded the selection ratio from the adjustment cell definitions used for Other Deductions and Cost of Goods.

Smoothing is not applicable when the donors within an adjustment cell are to be matched individually to the records being imputed. If cells contain too few donors, they must be combined to form cells of satisfactory size. An algorithm to collapse the adjustment cells so as to meet minimum size limitations can be automated readily, given a well specified set of rules.

To impute the magnitudes of change to Other Income, we collapsed adjustment cells to achieve a minimum of ten donors (with nonzero changes) per cell, subject to the following conditions. We collapsed first over the selection ratio class, combining adjacent cells sequentially until the number of donors reached or exceeded ten. If that was insufficient we collapsed over size. If both were insufficient we collapsed over the detailed industry class, but only within the same major industry.

For Other Income this collapsing strategy yielded 45 adjustment cells out of 345 possible cells. Among nonfinancial firms it was frequently necessary to collapse to the detailed industry level (implying generally fewer than 20 donors in such industry groups). Among firms in the banking industry, on the other hand, there was minimal collapsing.

5. MATCHING ON THE MAGNITUDE OF THE ORIGINAL AMOUNT

Within adjustment cells, as noted earlier, we performed a nearest neighbor match on the magnitude of the original amount to select the donor for each imputation. With this strategy, however, we were concerned about the potential for excessive matches to individual donors. To reduce the likelihood of this, we made provision for selecting the next nearest donor should the nearest donor have been used 20 times or more. If the next nearest neighbor was more than a specified distance from the record to be imputed, however, or had itself been used 20 times or more, we retained the initial selection--i.e., the nearest neighbor.

Despite the provision for selecting a secondary donor, matching to the nearest neighbor resulted in very uneven usage of the donors within a given adjustment cell. This was a consequence of differences in the distributions of original amounts between the small samples of donors and the records to be imputed in each adjustment cell. Donors in close proximity to other donors were used less often than donors more removed from their closest neighbors. The disparity in some adjustment cells was substantial.

Table 4 lists the logarithm of Other Income, the frequency of use, and the proportion of Other Income removed for each donor in the first adjustment cell, which corresponds to the first detailed industry class (see Appendix). The donors are listed sequentially by the original value of Other Income. The frequency of use among the 19 donors ranges from one to 20, with wide variation around a mean of 8.3 uses. It may be seen that the frequency of use is inversely related to the distance between adjacent donors, listed in the second column. The first and last donors also have high frequencies of use because they are nearest neighbors to all records to be imputed in the tails of the Other Income distribution.

Table 4

Donors Used to Impute Magnitudes of Change to Other Income Among Records in the First Minor Industry Class

Donor Number	Logarithm of OtherIncome	Distance between Donors	Frequency ofUse	Proportion of Other Income Removed
1	7.89	0.03	19	1.000
2	7.92	1.23	6	0.438
3	9.15	1.16	20	0.035
4	10.31	0.39	20	0.800
5	10.70	0.08	9	0.626
6	10.78	0.19	6	0.230
7	10.97	0.37	16	1.000
8	11.34	0.07	5	0.729
9	11.41	0.17	3	0.122
10	11.58	0.01	3	0.952
11	11.59	0.09	1	0.987
12	11.68	0.12	1	0.264
13	11.80	0.27	7	0.756
14	12.07	0.08	7	0.370
15	12.15	0.09	2	0.006
16	12.24	0.43	6	0.081
17	12.67	0.39	6	0.387
18	13.06	0.12	5	0.116
19	13.18		16	0.954
Mean			8.3	0.519

The proportion of Other Income removed from each donor displays broad variation around a mean of .519 (or .608 if weighted by the frequency of use). Adjacent donors often differ substantially in the proportion of Other Income removed--e.g., only .116 was removed from donor 18 while .954 was removed from the last donor. Consequently, the variation in frequency of use makes a large contribution to the imputation variance. In this adjustment cell the nearest neighbor match on the original amount of Other Income does prevent the imputation of implausibly small changes to donors with small original amounts, consistent with our intent. However, at the same time proportionately large changes are imputed to records with the largest original amounts, and this is contrary to our expectation.

Clearly this approach should be re-examined. Matching on the basis of rank order (percentile rank) within an adjustment cell would equalize donor use within the cell, but equal use is not the most favorable outcome if the distribution of original amounts differs markedly between the donors and the records to be imputed. Matching on the basis of **ranges** of original amounts may produce the optimal results. Such an approach allows donor use to vary with the broad shape of the distribution of original amounts among records to be imputed, but it does not link donor use too closely to fluctuations attributable to sampling.

Whatever the approach used to pair records within adjustment cells, the fact remains that the imputation of magnitudes could benefit greatly from the identification of stronger covariates. The within cell variance evident in Table 4 in the proportion removed from Other Income is substantial, and it increases the sensitivity of the imputed magnitudes to the distribution of donors within each cell.

6. CONCLUSION

Methods to reduce the cost of producing analytical data bases from administrative records have a clear need in the current budgetary environment. To lower the cost of one component of its corporation statistics program, the IRS has employed double sampling and imputed the unobserved edits. The resulting micro data can support extensive tabulations for detailed subpopulations. The production schedule and the need for several imputations per record favor an imputation methodology based upon the substitution of values from edited donors to unedited records. However, if subsampling is to produce substantial savings in editing costs, the resulting donors can be only a fraction of the total sample size. Even with data bases comprising near 100,000 records the potential donors may be small in number relative to what is required to make the best use of traditional methods of substitution-based imputation.

The problems posed by scarce donors were addressed in the revised imputation procedures employed by IRS in producing the 1985 corporation statistics of income. A key feature was the separate imputation of the change/no change outcome and the conditional magnitude of the change. The former could incorporate elements of model-based imputation while continuing to lend itself to automation. The results of the imputation of conditional magnitudes affirm the difficulty of performing imputations when donors are scarce. Further revisions are needed and could profit from additional research to identify stronger covariates of the conditional magnitudes of edited changes.

ACKNOWLEDGMENTS

This research was conducted by Mathematica Policy Research, Inc. (MPR) under contract to the Statistics of Income Division of the U.S. Internal Revenue Service. I am grateful to the SOI Division for its support, and I wish to thank Fritz Scheuren and Susan Hinkins in particular for allowing me to build upon their earlier work and for providing many helpful suggestions along the way. I would also like to acknowledge the efforts of David Edson and Cavan Capps of MPR in writing the imputation software and generating the tabulations reported here, and of Donald Rubin and Roderick Little in contributing to the development of the 1985 imputation procedures.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley and Sons, Inc.

Czajka, J.L. (1986). Imputation of selected items in corporate tax data: improving upon the earlier hot deck. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Czajka, J.L. (1987). Predicting edit outcomes: the strategic use of imputation in estimating corporate income statistics. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Hinkins, S. (1983). Matrix sampling and the related imputation of corporate income tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Hinkins, S. (1984). Matrix sampling and the effects of using hot deck imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Hinkins, S., and Scheuren, F. (1986). Hot deck imputation procedure applied to a double sampling design. *Survey Methodology*, 12 (December), 181-196.

Appendix
Industrial Classification

Wholesale Trade

1. Groceries and related; motor vehicles and automotive; furniture and home funishings; sports and recreation; drugs and sundries; apparel, piece goods and notions; alcoholic beverages
2. Machinery, equipment and supplies
3. Lumber and construction materials; metals and minerals; electrical goods; hardware, plumbing, heating equipment and supplies; other durable goods; paper and paper products; farm-product raw materials; chemicals, petroleum and allied products; miscellaneous nondurables

Retail Trade

4. Building, hardware and garden dealers; general merchandise and food stores
5. Automotive dealers and service stations
6. Apparel and accessory stores; furniture and home furnishings stores
7. Eating and drinking places
8. Miscellaneous retail stores; wholesale and retail trade not allocable

Manufacturing

9. Food and kindred products; tobacco manufacturers; miscellaneous manufacturing; nature of business not allocable
10. Textile mill products; apparel and other textile products; lumber and wood products
11. Furniture and fixtures; paper and allied products; primary metal industries; machinery, except electrical; motor vehicles and transportation equipment; instruments and related products
12. Fabricated metal products; electrical and electronic equipment

Services

13. Hotels and other lodging places; personal services; auto and miscellaneous repair services; amusement and recreation services
14. Business services; medical services; architectural and engineering services; accounting, auditing, and bookkeeping services; miscellaneous services
15. Advertising; legal, educational and social services; membership organizations

Other Nonfinancial Industries

16. Agriculture, forestry and fishing; mining; general building contractors, operative builders, and heavy construction contractors
17. Special trade contractors
18. Transportation and public utilities

Banking, Credit and Finance

19. Banks
20. Credit agencies; security, commodity brokers and services

Other Financial Industries

21. Insurance; insurance agents, brokers and services
22. Real estate operators and lessors of buildings
23. Other real estate; holding and investment companies

SESSION VII: CONTRIBUTED PAPERS

Chairperson: Daniel Kazprzyk, U.S. Bureau of the Census

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

RELATIONSHIPS OF MURDER CHARACTERISTICS TO TRIAL OUTCOMES AND TO CAPITAL PUNISHMENT: CANADA, 1961-1983

JANE F. GENTLEMAN and PAUL B. REED[1]

ABSTRACT

Statistics Canada collects data for all homicides (murders, manslaughters, and infanticides) in Canada known to the police, including information about the circumstances of the homicide, the characteristics of the persons involved, and the activities of the judicial system in response to the homicide. This detailed data set is compiled mostly from administrative records. In this paper, a few results are described from a study of murders which occurred during the period 1961-1983. Logistic regression models were fitted to analyze (1) the influence on the trial outcome of the characteristics of the murder and of the abolition of capital punishment in Canada in 1976; and (2) murder characteristics which are associated with the abolition of capital punishment.

1. INTRODUCTION

The data set on which this study was based contains comprehensive records of all homicides (murders, manslaughters, and infanticides) in Canada known to the police from 1961 through 1983, including information about the nature and circumstances of the homicide, the characteristics and relationships of the persons involved, and descriptions of the charges laid and the legal outcome of the suspect's trial. This detailed statistical file was produced mostly from administrative records. The data were collected by Statistics Canada as part of the Homicide Program of the Canadian Centre for Justice Statistics (CCJS). This program measures a phenomenon that can be enumerated **only** via administrative records. Homicides represent a relatively small population of events, and they entail exceptionally reprehensible behaviour, such that their detection and description are given a high order of importance and public action. The consequence of these two combined properties - the small population of events and the strategic importance accorded them - is that Statistics Canada's Homicide Program provides a census of known homicide events which is carried out with an extremely low level of coverage error. This demonstrates the potential that administrative data hold for generating a continuing longitudinal microdata set containing detail down to extremely fine geographic and temporal units.

The data set is organized by "incident," a single event in which homicide is committed, regardless of the number of suspects or victims involved. Within each incident, all victims and all known suspects are described. A total of 9642 homicide incidents, including data for 10627 victims and 9954 suspects, is on the data file used for this study.

[1] Jane F. Gentleman and Paul B. Reed, Analytical Studies Branch, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

During each month, the CCJS receives from police departments all over the country approximately 1500 copies of a crime report known as Form "C" - Crime Statistics (a blank copy of which is reproduced in Statistics Canada, 1986a). On this form (or equivalently on magnetic tape) is reported aggregate information about incidents involving homicide, sexual offences, assaults, robbery, prostitution, illegal drug activity, etc. For each homicide incident identified on Form "C", the Homicide Program of the CCJS receives from the appropriate police department a Homicide Return containing detailed information about the incident. A copy of a blank Homicide Return is reproduced in Statistics Canada (1986b). Coders at Statistics Canada use these data and other sources of information, such as newspaper clippings and telephone calls to police officials, to complete a Victim and Offence form for each victim and a Suspect form for each suspect. Data are then transferred from these forms to a magnetic tape containing historic homicide data, which was the source of the data for this study.

Currently under discussion at Statistics Canada are proposed changes to the Homicide Return and to some of the data code definitions (see Nabata, 1986, and Gentleman and Dixon, 1985, 1986, and 1987). Periodic review and updating of such a data collection system are necessary for a number of reasons. Over time, there are inevitable changes in the nature and availability of the data, and in the need and capability for utilizing them. In amassing homicide data, CCJS personnel must make compromises among a number of sometimes conflicting limitations and demands: controlling the cost of producing the data, balancing the response burden to police with the needs of data users, maintaining continuity in variable definitions while updating concepts and correcting problems, and ensuring confidentiality while supplying detailed data sets to the public. When scrutinized closely, all large data sets are found to have some problems. By analyzing Statistics Canada data "in house", the authors were able to provide feedback about possible improvements to the data, while their own analysis benefitted from their having access to microdata files and proximity to the data producers.

All active homicide incidents are monitored by the CCJS. The files are modified as necessary to bring them up to date by incorporating information from court records of the legal proceedings and legal status of the incidents, as well as data provided by Correctional Service Canada. The data file used for the present study was updated in 1986. Homicide incidents occurring later than 1983 were not studied, since information about the trial outcome for many of them was judged likely to be incomplete. A suspect for whom there was not at least one completed trial outcome was dropped from the analysis. This would include, for example, a suspect who was awaiting trial, or who died before conviction or acquittal. When there were multiple trials, the last trial on the records was used to obtain information about sentencing and conviction.

The homicide data base does not contain microdata for incidents of manslaughter and infanticide which occurred prior to 1974, because the time series was initially meant to describe only murder incidents (and manslaughter and infanticide are not defined as murder). Therefore, to avoid artificially inflating the homicide rate during 1974-1983, all suspects convicted of manslaughter or infanticide were excluded from this study. Also excluded were all juveniles tried in juvenile court. (All suspects - juvenile or adult - tried for murder in adult court were included.)

Table 1 lists the variables used in this study. These included two time-reference variables: Date of Disposition (which is the date of the suspect's acquittal or conviction), and Date of Offence. Two indicator variables were defined for the Existence of Capital Punishment, one indicating whether capital punishment existed at the Date of Disposition, and the other whether it existed at the Date of Offence. Two Rate variables were used: the Average Monthly Murder Victim Rate Over Last Year at Date of Disposition, and the Average Monthly Murder Victim Rate Over Last Year at Date of Offence.

Table 1

Variables Describing Homicide Incident

RE VICTIM
- Sex
- Age Group
- Marital Status
- Race

RE SUSPECT
- Sex
- Age Group
- Marital Status
- Race
- Education
- Whether or not Convicted
- Severity of Sentence
- Date of Disposition

OTHER
- Relationship of Suspect to Victim
- Means of Offence
- Apparent Motive
- Extra Circumstances
- Local Site of Offence
- Geographical Region
- Date of Offence
- Existence of Capital Punishment
- Murder Rate Over Last Year

For each discrete variable - i.e., for all variables except the Date and Rate variables - separate dummy variables corresponding to each outcome were defined for use as independent variables in logistic regressions (with one outcome - the "reference category" -omitted for each variable to avoid singularity of the design matrix).

The standard logistic regression model predicts $P = Pr(Y=1)$ to be $1/[1+\exp(-\hat{\alpha}-X\hat{\beta})]$, where Y is the dependent variable (whose observed value is either 0 or 1), X is a row vector of values of independent variables, $\hat{\alpha}$ is the fitted intercept, and $\hat{\beta}$ is the column vector of fitted regression coefficients of X. This is mathematically equivalent to assuming that the natural logarithm of the odds of Y being equal to 1 is a linear function of the independent variables: $\ln[P/(1-P)] = \alpha + X\beta$. In this paper, regression results are given in terms of the odds ratio $\exp(\hat{\beta}_j)$ where $\hat{\beta}_j$ is the fitted coefficient of a dummy variable. The odds ratio (OR) is the factor by which the estimated odds $\hat{P}/(1-\hat{P})$ are multiplied when the dummy variable changes from 0 to 1.

The computer program used to perform the logistic regressions of this study is described in SAS (1983). After fitting a full model, an efficient "fast backward" elimination procedure was used to select variables for a final "reduced" model, using an algorithm based on Lawless and Singhal (1978). Variables other than the intercept were deleted one at a time until no variable remaining in the model had a significance level above .05. (Variables could then re-enter the model if, after doing so, they had a significance level below .05.) The overall fit of a logistic regression was tested using the usual chi-square approximations to the likelihood ratio statistic and the score statistic. The Fraction of Concordant Pairs (FCP) was calculated for each regression as another

measure of the overall predictive capability of the fitted model. As calculated by the program in SAS (1983), the FCP is the proportion of correctly-ordered pairs of predicted values of the dependent variable, i.e., pairs which are ordered the same as the corresponding observed values.

Selected results from two types of logistic regression are described in this paper. Section 2 describes a logistic regression in which the dependent variable was Whether or not Convicted, and the independent variables were derived from all of the other variables listed in Table 1 except Severity of Sentence. Section 3 describes a logistic regression in which the dependent variable was Existence of Capital Punishment, and the independent variables were derived from all of the other variables listed in Table 1 except Date of Offence and Murder Rate Over Last Year. These results will be reported in greater detail elsewhere.

2. REGRESSION PREDICTING PROBABILITY OF CONVICTION

To determine which types of murders were associated with changes in the probability of conviction, and whether there were any effects of Existence of Capital Punishment (ECP), Date, and Rate, logistic regressions were performed using Whether or not Convicted as the dependent variable. Similarly, logistic regressions were performed using a polychotomous dependent variable Severity of Sentence. In order to reflect the behavior of the courts, the time-dependent variables in these regressions (ECP, Date, and Rate) were defined as of the Date of Disposition. A regression using Whether or not Convicted as the dependent variable might be interpreted as predicting the probability of a jury convicting a suspect. The polychotomous regression predicts the severity of the sentence (given conviction), which is determined by the judge, but which is limited to the range of sentences permitted by the specific type of conviction chosen by the jury. The judge also takes into consideration pre-sentence reports, and recommendations from the Crown, the jury, and the defence. Thus, the results of polychotomous regression reflect the judge's behavior, conditioned by these other influences.

Effects of the attitude of the judge and/or jury may be confounded with effects of the behavior of the police in laying charges. For example, if for some reason police laid a disproportionately large number of murder charges against innocent males, juries who acquitted them would appear to be acquitting a disproportionately large number of males. Similarly, the behavior of judges may be confounded with the behavior of suspects. For example, if killers who are male tend to commit more heavily punishable types of homicides than do killers who are female, an unprejudiced judge would tend to sentence males more heavily than females.

Table 2 shows results from a regression using Whether or not Convicted as the dependent variable. Odds ratios are given for selected independent variables in a full regression model. Both of the Sex variables were very highly significant (P=.00 to two decimal places). Male suspects (as compared to female suspects) had increased odds of conviction by a factor of more than 1.8, while suspects who were accused of killing a male victim (rather than a female one) had reduced odds of conviction by a factor of .65. The four combinations of Sex of Suspect and Sex of Victim were ranked as follows in descending order of odds of conviction: (1) a male suspect who is accused of killing a female victim (highest odds of conviction); (2) a male who is accused of killing a male; (3) a female who is accused of killing a female; (4) a female who is accused of killing a male (lowest odds of conviction). Another regression model which included an interaction term for Sex of Suspect with Sex of Victim showed the same results.

Table 2

Selected Logistic Regression Results
Dependent variable = Whether or not Convicted
Number of observations (number of suspects) = 6350

INDEPENDENT VARIABLE	ODDS RATIO	P-VALUE
Sex of Suspect = Female	1.00	---
Sex of Suspect = Male	1.88	.00
Sex of Victim = Female	1.00	---
Sex of Victim = Male	.65	.00
Race of Suspect = Caucasian	1.00	---
Race of Suspect = Native Canadian	1.34	.03
Race of Victim = Caucasian	1.00	---
Race of Victim = Native Canadian	.78	.07
Relationship of Suspect to Victim		
Victim = Suspect's Spouse	1.00	---
Victim = Suspect's Parent	.39	.00
Victim = Suspect's Child	1.50	.06
Existence of Capital Punishment (ECP)		
ECP = Not in Effect	1.00	---
ECP = In Effect	.71	.00

Native Canadian suspects had significantly higher odds of conviction (OR=1.34, P=.03) than Caucasian suspects.

Suspects accused of killing Native Canadians had a somewhat lower odds of conviction (OR=.78) than those accused of killing Caucasians. Although its P-value (.07) was large in the full regression model, this independent variable remained in the reduced model in which it achieved an odds ratio of .76 and a P-value of .04.

The reference group for the Relationship variable is suspects accused of killing their spouse. Suspects who were accused of killing their parent had the lowest odds of conviction (OR=.39, P=.00), while suspects accused of killing their child had the highest (OR=1.50). Although its P-value was high (.06), the latter independent variable remained in the reduced model with an odds ratio of 1.66 and a P-value of .00.

In predicting odds of conviction, Existence of Capital Punishment (ECP) was dominant among the three time-related variables (ECP, Date, and Rate). ECP was the only significant variable of the three (OR=.71, P=.00); Date and Rate had high significance levels in the full model (.28 and .47, respectively), and both were absent from the reduced model. These results indicate that whether or not an adjustment is made for Date and Rate, the odds of conviction were significantly lower when capital punishment was in effect than when it was not. This might be interpreted as indicating a reluctance to convict on the part of juries when capital punishment was at least a theoretical possibility.

3. REGRESSION PREDICTING EXISTENCE OF CAPITAL PUNISHMENT

The possibility of a relationship between the existence of capital punishment and the incidence of homicide has been investigated extensively over the past 30 years. There has been a rare consistency in the conclusions supported by this large corpus of work; persuasive and reliable evidence of capital punishment exerting a deterrent effect has not been found. Nevertheless, the effectiveness of capital punishment as a deterrent to murder is still extensively debated publicly. The intent of this study was to examine Canadian homicide data objectively to determine which if any murder characteristics are associated with the presence or absence of capital punishment. To this end, logistic regressions were performed using the dichotomous variable Existence of Capital Punishment as the dependent variable. In order to reflect the behavior of the suspect, Existence of Capital Punishment was defined as of the Date of Offence: if the Date of Offence was before July 1976, ECP equals one; otherwise, ECP equals zero. These regressions thus relate the suspect's characteristics and behavior to the date of the murder.

An important caveat is necessary at this point. The abolition of capital punishment in July 1976 occurred at the end of a period of numerous revisions of the definition and classification of murder, and of its associated penalty (see the historical summary in Appendix V of Statistics Canada, 1986b), with both **de jure** and **de facto** moratoria following the last execution in 1962; it also occurred during a period when other significant changes were being made in Canadian criminal law, some of which, such as the Gun Control Laws enacted in January 1978, would have quite intentionally had an impact on murder behavior and the justice system. Thus, it is a simplification to use July 1976 as an abrupt cutoff point for the existence of capital punishment, and the contributions of various factors may be confounded. Furthermore, the limitations of analyzing data from an uncontrolled experiment constrain us to phrase conclusions in terms of associations among variables, rather than in terms of causes and effects.

Table 3 shows selected results from a full regression model using Existence of Capital Punishment as the dependent variable. There was a significantly lower association of stabbing murders (OR=.59, P=.00) with ECP, compared to the reference category of shooting murders.

The variable Extra Circumstances is derived from a narrative section of the Homicide Return. From the information supplied by police, Statistics Canada coders record one of four possible outcomes -Drinking, Drugs, Gangland, or No extra circumstances mentioned. These outcomes are not mutually exclusive, and there are other problems in interpreting this variable, so the results should be considered uncertain. According to Table 3, alcohol-related murders were very much more associated with ECP than were other categories of Extra Circumstances (OR ranging from .47 to .64, and all P-values = .00).

The predicted odds that the murder occurred during the existence of capital punishment were not significantly different according to either the Sex of the Suspect or the Sex of the Victim.

The Apparent Motive variable is a subjective guess by the investigating police who, presuming that the suspect is guilty, indicate the reason they believe the suspect killed the victim. Compared to the reference category (anger, hate, argument, or quarrel), all other categories of Apparent Motive had higher odds ratios, most of them significant. Homicides apparently committed in self defence had the highest odds ratios (OR=3.52, P=.00).

Table 3 can be examined from the point of view of those who claim that capital punishment deters murders, and that it deters premeditated murders more than spontaneous ones. It is not a simple matter to determine whether and to what degree a murder was premeditated, but a few of the types of murders listed in Table 3 can probably

be categorized as relatively premeditated or spontaneous. To be consistent with a theory of deterrence, a spontaneous murder should have a higher odds ratio than a premeditated one, because a deterred murder would be less closely associated with the existence of capital punishment. It might be argued that Table 3 supports this theory for murders involving alcohol (as compared to those not involving alcohol) and for murders in which the apparent motive was self defence (as compared, perhaps, to those committed for purposes of revenge), if alcohol-related murders and murders committed in self defense are considered relatively spontaneous. This cannot be argued in the case of murders in which the means of offence was a knife (as compared to those in which a gun was used), if knife murders are considered more spontaneous than gun murders. However, it has been suggested by Scarff (1983) that firearms were to some degree displaced by knives as a Means of Offence as a result of the 1978 Gun Control Laws.

Table 3

Selected Logistic Regression Results
Dependent variable = Existence of Capital Punishment
Number of observations (number of suspects) = 6369

INDEPENDENT VARIABLE	ODDS RATIO	P-VALUE
Means of Offence = Gun	1.00	---
Means of Offence = Knife	.59	.00
Extra Circumstances =		
Alcohol	1.00	---
Not Alcohol	.47-.64	.00
Sex of Suspect = Female	1.00	---
Sex of Suspect = Male	NS	NS
Sex of Victim = Female	1.00	---
Sex of Victim = Male	NS	NS
Apparent Motive =		
Anger, Hatred, Argument, Quarrel	1.00	---
Sex Assault, Robbery, Theft, Other Crime	NS	NS
Revenge	1.37	.00
Jealousy	2.20	.00
Unknown	2.20	.00
Mentally Ill	2.23	.00
Other	2.57	.00
Self Defence	3.52	.00

NS = Odds ratio not significantly different from 1.00 at .05 level.

REFERENCES

Gentleman, Jane F., and Dixon, Daniel P. (1985). Assessment of the Analytic Usefulness of Statistics Canada Homicide Data. Internal Statistics Canada report. Social and Economic Studies Division.

Gentleman, Jane F., and Dixon, Daniel P. (1986). Comments on Revised Coding Instructions for Statistics Canada Homicide Date. Internal Statistics Canada report. Social and Economic Studies Division.

Gentleman, Jane. F., and Dixon, Daniel P. (1987). Comments on Proposed Changes to the Homicide Return and to the Homicide Data Coding System (1987). Internal Statistics Canada report. Social and Economic Studies Division.

Lawless, J.F., and Singhal, K. (1978). Efficient Screening of Nonnormal Regression Models. Biometrics, 34, 318-327.

Nabata, Tony (1986). The Homicide Development Project - The Homicide Return. Internal Statistics Canada report. Canadian Centre for Justice Statistics.

SAS Institute Inc. (1983). SUGI Supplemental Library User's Guide, 1983 Edition. (Description of the LOGIST Procedure by Frank E. Harrell, Jr.) SAS Institute Inc., North Carolina, 181-202.

Scarff, Elisabeth (1983). *Evaluation of the Canadian Gun Control Legislation. Final Report*. Solicitor General Canada.

Statistics Canada (1986a). *Canadian crime statistics*, 1985. Catalogue 85-205 Annual. Canadian Centre for Justice Statistics.

Statistics Canada (1986b). *Homicide in Canada, 1984, A Statistical Perspective*. Catalogue 85-209 Annual (entitled *Homicide Statistics* for 1981 data and earlier). Canadian Centre for Justice Statistics.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

THE USE OF ADMINISTRATIVE RECORDS IN CANADA FOR ESTIMATING POPULATION AND COMPONENTS OF POPULATION CHANGE

RAVI B.P. VERMA and RONALD RABY[1]

ABSTRACT

The adequacy of the Family Allowance and Revenue Canada tax files for providing the estimates of emigrants from Canada and the interprovincial migration data was examined. Their applications in estimating total population for Canada, provinces and territories, were evaluated with respect to the 1986 Census counts. It was found that these two administrative files provided consistent and reasonably accurate series of data on emigration and interprovincial migration during the period 1981-86. The estimate of emigrants from the Family Allowance file could be improved by using the ratio of adult to child emigrant rates computed from the immigration file.

1. INTRODUCTION

Population estimates for Canada, provinces and territories, census divisions, and census metropolitan areas are based on the latest census counts and several administrative data sources; Revenue Canada tax and Family Allowance files for migration, Vital Statistics registration for births and deaths, and Immigrant Visa and Record of Landing Registration for immigration. The strengths and weaknesses of these administrative files in estimating population and migration in comparison with the 1981 Census data were discussed elsewhere. (Statistics Canada, Catalogue 91-528E, 1987; Verma and Parent, 1985; and Norris, Britton and Verma, 1982). In this paper, the accuracy of the Family Allowance and Revenue Canada data sources in estimating population for provinces and territories will be highlighted with reference to the 1986 Census counts. The performance of these administrative files in 1986 will be assessed in terms of comparisons based on 1971, 1976 and 1981 data.

2. DATA SOURCES AND THE METHODS OF ESTIMATION

Since family allowance and tax files do not cover the whole universe, some emigrants and interprovincial migrants need to be estimated. This section describes the procedures for estimating i) internal migration, ii) emigrantion and iii) total population.

i) Internal Migration

Two administrative files are used to produce annual and quarterly estimates of interprovincial migration. Preliminary estimates are derived from Family Allowance

[1] Ravi B.P. Verma and Ronald Raby, Demography Division, Statistics Canada, 4-A Jean Talon Building, Ottawa, Ontario. K1A 0T6

files. Final data on interprovincial migrants are estimated from Revenue Canada income tax files.

Preliminary Estimates

Recipients of family allowance cheques must notify the department of Health and Welfare of any changes in address. From these files, the movements of recipients are compiled monthly, by province of origin and destination and, in terms of the number of families by size (the number of children per family receiving the allowance). Coverage of the population by family allowance is comparable to that of the census (Statistics Canada, Catalogue 91-528E: 46). The number of adult migrants is estimated by using both the child migration figures derived from Family Allowance files and the ratios of adult out-migration rates to child out-migration rates ($f_{j,k}$) based on the most recent Revenue Canada tax file calculated for a 1 or 2 year period prior to the estimate reference date. Estimates of the number of adult interprovincial out-migrants (aged 18+), and for all age groups are calculated as follows:

$$\hat{M}_{(j,k),18+} = \frac{M_{(j,k),0-17}}{P_{j,0-17}} \cdot f_{(j,k)} \cdot P_{j,18+} \tag{1}$$

$$f_{(j,k)} = \frac{M'_{(j,k),\ 18+}}{\hat{P}_{j,18+}} \div \frac{M'_{(j,k),\ 0-17}}{\hat{P}_{j,0-17}} \tag{2}$$

$$\hat{M}_{(j,k),0+} = \hat{M}_{(j,k),18+} + M_{(j,k),0-17} \tag{3}$$

where:

$\hat{M}_{(j,k),0+}$ = estimated total number of persons out-migrating from province j to province k

$\hat{M}_{(j,k),18+}$ = estimated number of adult out-migrants (aged 18+) from province j to province k

$M'_{(j,k),18+}$ = Number of adult out-migrants from province j to province k derived from Revenue Canada tax file

$M'_{(j,k),0-17}$ = Number of child out-migrants (aged 0-17) from province j to province k derived from Revenue Canada tax file

$M_{(j,k),0-17}$ = the number of child out-migrants from province j to province k, based on Family Allowance files

$P_{j,18+}$ = the estimated number of adults in province j, computed residually from the total population estimates (Demography Division) and the estimates of the child population based on Family Allowance counts

$P_{j,0-17}$ = the total number of children receiving family allowance payments in province j

$f_{(j,k)}$ = estimation factor for adult migrants from province of origin j to province of destination k, based on estimates of migration data from Revenue Canada tax file.

$\hat{P}_{j,18+}$ = number of adults in province j; Demography Division Population Estimate

$\hat{P}_{j,0-17}$ = number of children in province j; Demography Division Population Estimate

Final Estimates

Revenue Canada tax files are used to produce final estimates of interprovincial migrants. All individuals receiving an annual income above a specified minimum are required to file an income tax return by the end of April of each year. Migrant tax filers are identified by a comparison of current and previous areas of residence according to two consecutive tax returns. Information on the characteristics of dependents is obtained by imputation; the number and ages of the dependents are determined from the total amount of the personal exemptions of the filer. An adjustment is made for the population not covered by the Revenue Canada system; this includes people who neither file an income tax return nor appear as dependents in another filer's return (Norris and Standish, 1983; Statistics Canada, Catalogue 91-528).

ii) Emigration

In Canada, there is no system of recording emigrants; therefore, their numbers need to be estimated. Emigrants can be identified from Revenue Canada income tax files with an "out-of-Canada" address one year, and an "in-Canada" address for the previous year. Also the emigrant status of children under 17 years of age is compiled directly from change of address notifications for family allowance recipients. By combining information from these two administrative files, both preliminary and final estimates of emigrants are generated. The estimation procedures are similar to those used to estimate preliminary inter-provincial migration.

$$E_j = \left[\frac{E_{j,0-17}}{P_{j,0-17}} \cdot F_c \cdot P_{j,18+} \right] + E_{j,0-17} \tag{4}$$

$$F_c = \frac{\hat{E}_{c,18+}}{\hat{P}_{c,18+}} \div \frac{\hat{E}_{c,0-17}}{\hat{P}_{c,0-17}} \tag{5}$$

$$E_c = \sum_{j=1}^{12} \left[E_j \right] \tag{6}$$

where:

E_j = estimated annual number of emigrants from province j

E_c = estimated annual number of emigrants from Canada

$E_{j,0-17}$ = the number emigrants from province j aged 0 to 17 who are eligible for family allowance

$P_{j,0-17}$ = the number of children in province j who are eligible for family allowance

$P_{j,18+}$ = adult population of province j obtained by subtracting the number of children eligible for family allowance from the total estimated population

F_c = annual adjustment factor for estimating adult emigration from Canada, based on Revenue Canada tax files.

$\hat{E}_{c,18+}$ and $\hat{E}_{c,0-17}$ = estimated numbers of adult and child emigrants from Canada, based on Revenue Canada tax files.

$\hat{P}_{c,18+}$ and $\hat{P}_{c,0-17}$ = estimated June 1st population of adults and children of Canada, Demography Division.

iii) Total Population

Quarterly and annual estimates of the total population of Canada, provinces and territories, and the annual totals for census divisions and census metropolitan areas are produced by the component method. At the national level, the number of births and immigrants are added to, and the number of deaths and emigrants are subtracted from the base population (taken from the latest Census of Canada). At the provincial level and for local areas, data on estimates of internal migration are also taken into account.

3. EVALUATION OF THE ADMINISTRATIVE DATA ON THE COMPONENTS OF POPULATION CHANGE

Each of the components of population change (births, deaths, immigrants, and estimates of emigrants and interprovincial migrants) may contain a degree of bias and error. However, the data on births, deaths and even immigration can be regarded as being fairly accurate, while the estimates of emigrants and interprovincial migrants are less so. The methods of estimation of emigrants and internal migration were thoroughly updated in 1982 (see Statistics Canada, Catalogue 91-528E). They are evaluated below.

Emigration Data

Table 1 presents a comparison of estimates of emigrants from Canada by different methods and data sources for the periods 1976-81 and 1981-86. During the period 1981-86 the estimate of emigrants by the residual method, subtracting the intercensal change of census counts between 1981 and 1986 (unadjusted for census undercoverage) from the natural increase and immigration, is considerably higher than the estimate based on the family allowance method. Since the births, deaths and immigration data are assumed to be accurate, the higher estimate of emigration by the residual method can be attributed to the differential in the undercoverage rates for 1981 and 1986. On adjusting the 1981 and 1986 Census counts for the undercoverage rates, 2.01% and 3.21% respectively, the estimate of emigrants by the residual method was found to be 134,857. This estimate is lower than that from the family allowance method, (235,481), and also from the Revenue Canada Tax files (165,272).

Thus, the lower estimated number of emigrants at the national level by the residual method, using the census data adjusted for undercoverage was possibly due to the fact that the element of the differential overcoverage of the rates in the 1981 and 1986 Censuses are not taken into account. No estimate of the rate of overcoverage is calculated from the Reverse Record Check study, but it can be assumed to be similar to the one observed in the U.S.A. and to be 25% of the undercoverage rate. On adjusting the 1981 and 1986 Census counts for the net coverage rates by 1.51% and 2.40%, respectively,

the estimate of the emigrants was close to that from the family allowance based estimate, 218,148 vs. 235,481.

Table 1

Estimates of Emigrants by Different Methods, Canada, 1976-81 and 1981-86

Method	1976-81	1981-86
Residual*		
(a) Unadjusted	277,558	476,373
(b) Adjusted for Undercoverage	196,955 (1)	134,857 (1)
(c) Adjusted for Net Undercoverage	194,155 (2)	218,148 (2)
Revenue Canada Tax File**	207,420	165,272
Family Allowance Method	278,624	235,481
Reverse Record Check***	296,724	288,376

* Residual Method:
Emigrants = (Births - Deaths) + (Immigrants)) -Intercensal growth of population between t and t+5

* obtained directly, independently of Family Allowance Data. The estimated number emigrants is equal at $\hat{E}_{c,18+} + \hat{E}_{c,0-17}$ defined on page 4.

*** emigrants traced in the sample of the RRC, inflated to the total population level (see no. 2, July 1988).

(1) Adjusting the 1976 Census counts for undercoverage rate (2.04%), the 1981 Census counts for undercoverage rate (2.01%), and 1986 Census counts for rate of undercoverage (3.21%).

(2) Adjusting the 1976, 1981 and 1986 Census counts for the net undercoverage rates by 1.53%, 1.51% and 2.40%, respectively. The net undercoverage rates represent about 75% of the missed persons estimated from the Reverse Record Check.

Source: Demography Division, Statistics Canada

Estimates produced by using the same methods, for the period 1976-1981, don't confirm these conclusions. The numbers of emigrants estimated by the residual method adjusted for net undercoverage were 194,155, comparable to those based on Revenue Canada Tax File (207,420), but lower than emigrants estimated by the family allowance method (278,624) or the reverse record check (296,724).

One possible source of error in the current method is evaluated by the F_C factor, the adult-child emigrant ratios used to estimate the number of emigrants from Canada for the years 1981-86. These ratios are obtained from the Revenue Canada tax files data.

Table 2 shows the estimates of emigrants from Canada for different values of F_C (adult-child emigrant ratios) for the years 1981-86 based on different data sources. One can see that the F_C factors from the Revenue Canada tax files are less than unity whereas the ratios are higher than unity from three other data sources; interprovincial

migration data from income tax files, immigration files and the Canadian emigrants to the U.S.A. The numbers of emigrants from these sources are also estimated to be higher than the official number of emigrants (235,481).

Table 2

Estimates of Emigrants from Canada for Different Values of F_c (Adult-Child Emigrant Ratios), 1981-86

Data Source of F_c	Value of F_c Factor by Single Year					Number of Emigrants
	1981-82	1982-83	1983-84	1984-85	1985-86	1981-86
1. Revenue Canada Tax Files	0.8698	0.8768	0.9052	0.8592	0.8592	235,481
2. Interprovincial Migration Data from Income Tax Files*	1.0760	1.1000	1.0664	1.0290	1.0029	265,816
3. Immigration Data*	1.0801	1.0926	1.1723	1.1254	1.0694	275,762
4. Canadian Emigrants to the U.S.A.	1.2300	1.2774	1.3196	1.3745	1.4232	316,268

* On the hypothesis that the age structure of emigrants is similar to that of immigrants or interprovincial migrants.

Source: Demography Division, Statistics Canada

For each data source of the F_c factor there are some annual variations. In particular, the values of the F_c factor for Canadians emigrating to the United States are relatively high. This suggests that in comparison with child emigrants, 23% to 42% more adults and older retiree Canadians had emigrated to the U.S. This is not surprising, as the southern States of the U.S. have always been attractive locations for Canadian retirees. Hence, the value of the F_c factor based on U.S. data may not be suitable for estimating Canadian emigrants to countries other than the U.S.A.

Similarly, the values of the F_c factor from the interprovincial migration, based on the income tax file, suggest that the range of 0% to 10% more adult migrants have exceeded the child migrants over the years, 1981 to 1986. Among these adult migrants, there could be a greater proportion of younger adults. Hence, this source of data is also very specific and not suitable for computing the F_c factor.

Since, according to some authors (Beaujot and Rappak, 1988), emigrant flow data are associated with the immigrant flow data, one could compute an F_c factor from the immigration file. The values of the F_c factor from the immigration file are intermediate between the interprovincial immigrants and the U.S. emigrants. The official estimate of emigrants (235,481) for the period 1981-86 seemed to be underestimated in comparison with the figure based on the F_c factor from the immigration file (275,762). The latter estimate is close to that derived from the 1986 Reverse Record Check Study (288,376). By increasing the number of emigrants to 275,762, the 1986 error of closure between the population estimate and census counts, for Canada, is reduced from 0.95% to 0.79%.

In sum, the estimates of emigrants could be improved by taking the factor (F_c) of the adult emigrants to child emigrants ratio from the immigrant data of Canada Employment and Immigration instead of using the Revenue Canada tax files.

Interprovincial Migration Data

In order to test the accuracy of the estimates of interprovincial migration data from the Revenue Canada tax file two types of evaluation have been conducted: i) the consistency of the two sets of interprovincial migration data derived from the Revenue Canada tax and Family Allowance files; and (ii) a comparison of the error of closure of the population estimates based on the two sets of internal migration data with the 1986 Census counts.

Table 3 presents a comparison of net interprovincial migration estimates derived from three sources; the Revenue Canada tax files, Family Allowance file and the residual based net migration. For each province, the two estimates of internal migration data from the Revenue Canada tax and Family Allowance files are consistent in terms of the direction of the net migration. For both sets of data the same four provinces had positive net migration during the period 1981-86. In the other provinces net migration was negative.

Table 3

Estimates of Net Interprovincial Migration from Family Allowance Files, Income Tax Files and Residual Method,* Canada and Provinces, 1981-1986

Geographic Area	Family Allowance	Income Tax	Residual Estimates
CANADA	0	0	-238,178
Nfld.	-14,837	-15,051	-26,111
P.E.I.	293	751	-509
N.S.	5,204	6,895	-4,095
N.B.	-2,239	-65	-11,212
Qué.	-76,040	-81,254	-167,286
Ont.	115,497	121,767	57,147
Man.	-3,700	-2,634	-8,180
Sask.	-668	-2,974	-13,564
Alta.	-34,073	-31,676	-50,811
B.C.	13,289	7,382	-12,418
Yukon	-2,381	-2,775	-1,643
N.W.T.	-345	-366	504

* The residual method for estimating net interprovincial migration is given below:

Net Migration = - (Births - Deaths) + (Immigration - Emigration) + Growth of Census Population between time t and t+5

Source: Demography Division, Statistics Canada

The estimates of net interprovincial migration from Family Allowance and Revenue Canada tax files are not strictly comparable to those obtained by the residual method. By definition, in Canada, the sum of net interprovincial migration should be equal to zero.

However, the sum is about 238,178. (See Table 3). In addition, the differences between the residual and the Revenue Canada tax and Family Allowance based net interprovincial migration are considerably higher in five provinces: Newfoundland, New Brunswick, Quebec, Ontario, and Alberta.

The coefficient of variation (the ratio of the standard deviation of the average absolute error of closure among provinces to the average absolute error of closure) was used to measure the relative accuracy of the internal migration data, assuming that the other components of population change are accurate. Statistically, a coefficient of variation of 20% to 30% is considered acceptable.

Table 4 shows the coefficient of variation (computed from figures in Table 5) between population estimates based on two sets of internal migration data and the census counts for the years, 1971, 1976, 1981 and 1986. Prior to the 1976-81 period, the coefficients of variation for the migration data from tax files were 50% higher than those for the family allowance file. This observation could have been expected, as the methodology for estimating migration from tax files was in the developmental stage and thus the estimates of migration data were experimental. Futhermore, in estimating the number of interprovincial migrants the factor f_j (adult to child migration rates) was based on the Census mobility data. This approach was found to be less satisfactory than the current method of estimation. However, for the periods 1976-81 and 1981-86, the gap in the coefficient of variation between the tax and family allowance migration data has narrowed considerably.

Table 4

Coefficients of Variation of the Average Absolute Error of Closure between the Population Estimates and Census Counts among Provinces (n=10), by Source of Interprovincial Migration Estimates, 1966-1971, 1971-1976, 1976-1981 and 1981-1986

Period (t,t+5)	Source	AAE (t+5)	Standard Deviation	C.V.
		(1)	(2)	(3)=[(2)-(1)] x 100
1966-1971	Tax	0.91	0.2863	31
	FA	1.33	0.2642	20
1971-1976	Tax	0.44	0.1317	30
	FA	0.97	0.2135	22
1976-1981	Tax	0.69	0.2463	36
	FA	0.86	0.2855	33
1981-1986	Tax	1.07	0.1496	14
	FA	1.01	0.1570	16

Note: AAE : Average absolute error of closure
C.V. : Coefficient of variation
Tax : Interprovincial migration data source file (Revenue Canada Income Tax)
FA : Interprovincial migration data source file (Family Allowance File)

Source: Computed from Table 5.

In 1981, the coefficient of variation for the tax-based migration data was 9% higher, whereas for 1986 it was 19% lower than the coefficient of variation based on the family

allowance. These values are small. Hence, these two sets of migration data are highly comparable. They produce the provincial estimates and the errors of closure which have the same level of variation among provinces. Since the coefficient of variation is under 20% in 1986, each of these sources provide acceptable data on internal migration.

In conclusion, estimates of interprovincial migration from the Revenue Canada tax files for the period 1981-86 are consistent with those from the Family Allowance file. At the provincial level they yield less variation in the errors of closure, i.e., difference between the population estimates and census counts.

Table 5

Error of Closure Between Alternative Population Estimates and Census Counts by Province and Territory Using Two Sets of Interprovincial Migration Data, 1971, 1976, 1981 and 1986

	Percent Error of Closure(1)							
Geographic Area	**Tax 1971**	**FA**	**Tax 1976**	**FA**	**Tax 1981**	**FA**	**Tax 1986**	**FA**
Newfoundland	-2.08	-1.64	0.49	1.34	1.63	2.30	1.97	2.01
Prince Edward Island	-2.09	-2.01	0.17	2.11	-0.05	1.02	0.99	0.63
Nova Scotia	-1.68	-2.39	-0.20	1.18	0.30	0.40	1.24	1.04
New Brunswick	-1.93	-2.65	-1.29	1.81	0.13	0.54	1.58	1.04
Quebec	-0.33	-0.97	-0.05	-0.18	-0.30	-0.07	1.32	1.40
Ontario	0.11	0.99	0.15	0.16	0.64	0.37	0.72	0.65
Manitoba	0.29	0.38	-0.27	0.39	1.07	0.87	0.51	0.41
Saskatchewan	0.44	-0.33	0.45	0.37	-0.31	0.28	1.08	1.31
Alberta	-0.14	0.52	-1.07	-1.11	-2.39	-2.64	0.73	0.63
British Columbia	0.01	-1.34	0.28	-1.10	0.03	-0.07	0.59	0.79
Yukon	-5.36	-5.99	-0.87	3.79	-1.98	2.06	-4.78	-3.10
Northwest Territories	-2.12	2.64	-12.98	-3.39	-7.08	0.43	-1.44	-1.40
Average Absolute Error								
10 provinces	0.91	1.33	0.44	0.97	0.69	0.86	1.07	1.01
Provinces and Territories	1.38	1.82	1.52	1.41	1.33	0.92	1.41	1.22

Note: From 1976 to 1981, Revenue Canada data for children were available for age group (0-15) only. Therefore the $f_{(j,k)}$ factors were calculated using migrants aged (0-15) and 16+ instead of (0-17) and 18+.

(1) Error of closure is calculated by using the following equation:

$$\text{Error of closure} = \left(\frac{\text{Estimate} - \text{Census}}{\text{Census}}\right) \times 100$$

Source: Estimates of interprovincial migration based on Family Allowance data, (FA) Demography Division, Statistics Canada.
Estimates of interprovincial migration based on tax data, (Tax) Small Area and Administrative Development Division, Statistics Canada.

4. CONCLUSION AND DISCUSSION

Family Allowance and Revenue Canada tax files are playing important roles in providing a consistent series of emigration and internal migration data for Canada, provinces and territories. The estimates of emigrants and interprovincial migrants from these files during the period 1981-86 are acceptable for estimating total population.

At the national level, the 1986 error of closure (difference between the 1986 population estimates and census counts) was higher than for the preceding census years, 1971 to 1981. In addition, the differences between the population estimates and census counts among provinces in 1986 were positively biased, indicating that in all provinces the estimates were higher than their census counts. How can one explain these discrepancies?

The answer to this question is closely related to the coverage of the 1981 Census population which was used as the bench mark, and the 1986 Census population. The 1981 undercoverage rate from the Reverse Record Check for Canada was estimated to be 2.01%. The similar rate for the 1986 Census is considerably higher, 3.21%. Thus, census undercoverage is emerging as an important factor in determining the source of error of closure between the population estimates and the census counts.

REFERENCES

Beaujot, R. and Rappak, J.P. (1988). *Emigration from Canada: its importance and interpretation*. Ottawa: Employment and Immigration Canada.

Statistics Canada, Catalogue 91-528E (1987). *Population Estimation Methods, Canada*. Ottawa: Ministry of Supply and Services Canada.

Norris, D., Britton, M., and Verma, Ravi B.P. (1982). The Use of Administrative Records for Estimating Migration and Population. *Statistics of Income and Related Administrative Record Research: 1982*, Washington, D.C. Department of the Treasury, Internal Revenue Service.

Norris, D. and Standish, L.D. (1983). A Technical Report on the Development of Migration Data from Taxation Records. Technical Report. Ottawa: Small Area and Administrative Data Development Division, Statistics Canada.

Verma, Ravi B.P. and Parent, Pierre (1985). An Overview of the Strengths and Weaknesses of the Selected Administrative Data Files. *Survey Methodology*, Vol. 11, No. 2, pp. 171-179.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

STATISTICS ON ADMINISTRATIVE REGISTERS IN MEXICO PRESENT SITUATION AND PROBLEMATIC

MA. ELENA FIGUEROA M.[1]

ABSTRACT

In Mexico, the National Institute of Statistics, Geography and Data Processing (Instituto Nacional de Estadistica, Geografia e Informatica - INEGI) was created in 1983. Its main functions are, the coordination of activities related to the production of statistical information in Mexico. At the end of 1983 the Institute began its decentralization by creating Regional Directorates, located in 10 federal states.

Currently vital statistics are captured and processed regionally and national figures are obtained in central offices. The same procedure used for vital statistics is proposed for statistics on Health, Education and Culture, and Justice and Labour Relations. There exist several problems which must be solved: continuous update of Directories of Informants, continuous training of regional personnel in Information capture and coding, training on analysis and interpretation of information generated in regional offices, coordination with Social Institutions of decentralization of the information they produce, quality control, both of coverage and processing of data, efficient procedures for coordination with regional offices. These problems and future direction of developments will be discussed in this paper.

1. INTRODUCTION

Statistical information in Mexico has been compiled since 1882 by the Mexican Bureau of the Census and Statistics, at the present an agency of the National Institute of Statistics, Geography and Data Processing.

Although Mexico has faced a variety of problems regarding financial and human resources in order to develop statistical data, it has been possible to complete 10 population censuses, and to generate a wide variety of economic and social statistics from different administrative records.

Census operations have greatly improved as a result of changes introduced every 10 years using national and international experiences, data processing development, necessity of reliable updated data concerning dynamics of population, among others.

[1] MA. Elena Figueroa Marquez, National Institute of Statistics, Insurgentes Sur 795, 120. PISO, Colonia Napoles, C.P. 03810 Mexico

Production of statistics in Mexico has become extremely difficult and expensive due to the large area of the country, its geographical characteristics, high population growth rates and administrative problems of a centralized process of statistical production.

Administrative records had been neglected because of reasons and priority have been in census collection.

In order to promote statistical development at the national and regional level, in 1983 creation of ten Regional Offices throughout Mexico was started by the National Institute of Statistics, Geography and Data Processing (INEGI).

Each regional office covers an average of three States and - it's main responsibility is to bring data information services closer to local users and to speed up the data collection and compiling process.

Activities of each regional office are coordinated by central office in order to achieve consistency between national and regional projects.

Central Statistical Office compiles and processes data from Population and Economic Censuses, Economic Surveys and Administrative records in Demographic, social and economic areas.

Central Office is divided in five areas in order to fullfil those activities.
Such Areas are the following:

1. Censuses
2. Short Term Statistics
3. National Accounts and Economic Statistics
4. Demographic and Social Statistics
5. Technical Assistance.

Collection and data processing from administrative records is the responsibility of Demographic and Social Statistics. Demographic Statistics Division produces statistics on births, deaths, fetal deaths, marriages, divorces, migration and international tourism. Social Statistics Division is responsible for the following statistics:

1.	Cultural:	Cinemas, museums and public shows (theatres, - stadiums, etc.) data.
2.	Labor relations:	Labor conflicts and agreements, strike summons and strikes data.
3.	Health and Social Security:	Medical services (public and private) and social welfare data.
4.	Safety and Public order:	Delinquency and suicides data.

Generation of such statistics was started different years ago. Vital statistics are the oldest ones being compiled since 1893.

In 1983, an evaluation of statistical production process was carried out by Demographic and Social Statistics Division. The following problems were noted:

1. Delays on data delivery from different sources.
2. Out of date and incomplete data sources of "Directories".
3. Long periods of production due to centralization and heavy load of codification.

4. Out of date statistical questionnaires and their conceptual and operating issues for electronic processing of information.

5. Out of date processing programs with operating problems.

In order to solve some of the above mentioned problems it was decided to decentralize statistical production to regional offices, under central supervision using standard methods. Such process was initiated with vital statistics due to its volume and importance.

Two main areas are involved with decentralized production process. Their functions are the following:

1. Central Area - Responsible for regional production process normativity and National integration of data.

2. Regional Area - Responsible for manual and electronic process operation as well as for analysis - and spreading of data at state and regional levels.

Present demographic information process is shown in enclosed diagram.

Similar decentralization process will be held for social statistics. Different official institutions had been consulted in order to define data delivery procedure to the various regional offices.

In order to improve and to facilitate such decentralization will require further attention:

1. Continuous up to date of Directories of Informants.

2. Continuous training and motivation of personnel of regional offices particularly in aspects related with data collection and codification.

3. Training of personnel in regional offices for analysis and interpretation of statistical information.

4. Quality controls, both for coverage and processing.

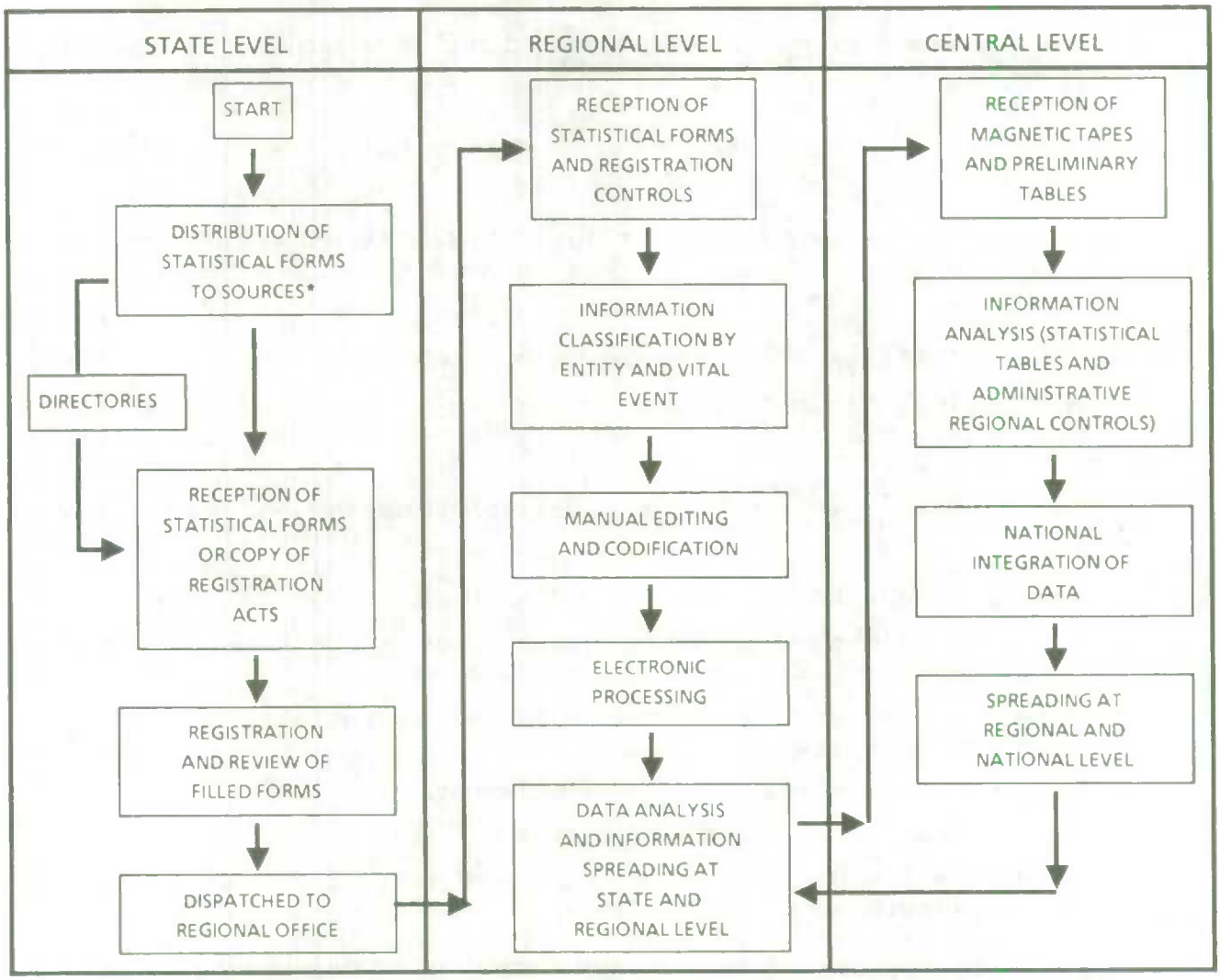
5. Efficient coordination procedures with regional offices.

6. Institutional coordination so as to guarantee effective - participation in updating and decentralization of social statistics.

In order to satisfy timely the demand for statistical information and at the same time achieve better opportunity in publication of data, Demographic and Social Statistics Area is interested in the development of statistical indicators and models which allow for calculation of estimates to satisfy such needs.

Same consideration is also given for the evaluation of the present statistical information and their linkage to other data sources as census and surveys.

DATA GENERATION PROCESS
FOR VITAL STATISTICS
- SINCE 1985 UP TO DATE -



* ONLY FOR DEATHS, FETAL DEATHS AND DIVORCES. BEGINNING 1988 DEATHS WILL BE RECEIVED IN OFFICIAL REGISTRATION ACTS AND CERTIFICATES.

SESSION VIII: CONTRIBUTED PAPERS

Chairperson: John Coombs, Statistics Canada

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

UPDATING TAX RETURN SELECTION PROBABILITIES IN THE CORPORATE STATISTICS OF INCOME PROGRAM

SUSAN HINKINS, HOMER JONES, and FRITZ SCHEUREN[1]

ABSTRACT

The Statistics of Income (SOI) sample of Corporate tax returns is essentially fixed in size at about 90,000 annually. Because of population shifts and overall growth, the selection probabilities typically are lowered each year. Issues raised by this process include the extent to which such adjustments affect cross-section estimates, estimates of short year-to-year change and longitudinal sample composition. The present paper begins with some background on the nature of the current corporate SOI sample design and how it has dealt with these issues historically. An evaluation is then conducted of alternative updating strategies and these are contrasted with current practice.

1. INTRODUCTION

Since 1951, the U.S. Internal Revenue Service has been sampling corporation tax returns to produce annual estimates of economic characteristic using tax data. Through the years, the process for collecting information and making estimates has evolved and changed as a result of shifts in the population of corporations and revisions in the tax law.

Advancements in computer and statistical technology have also brought about modifications to the procedures. Unfortunately, due to a number of practical limitations -- the constant growth of the population of organizations filing corporation tax returns, the limitation on the number of returns which we can process due to budgetary restrictions, and the short time period within which we are permitted to process the sampled returns -- sampling rates have gone down almost continuously over the years. These declining sampling rates make it more difficult to keep corporations in the sample over a period of years and, therefore, hamper accurate measures of change from year to year.

This paper focuses on recent modifications to the sample design that could improve estimates of year-to-year change. Some such design features are already in place, and other options are being considered for the future. The topics touched on briefly include several relevant features of the corporate design, description of the problem of estimating year-to-year change, and recent revisions to the sample design. Of course, as we consider such modifications, we must also look at possible design effects on cross-sectional estimates. These are examined and some thoughts on future directions are raised in the concluding section.

[1] Susan Hinkins, Homer Jones, and Fritz Scheuren, Statistics of Income Division, Internal Revenue Service, 1111 Constitution Ave., NW, Washington, DC 20224, U.S.A.

2. BACKGROUND

The primary objective of the corporate sample is to derive accurate annual estimates of economic variables. For example, recent cross-sectional estimates have included - -

- total assets for all 1984 corporations ($11.1 trillion);
- total receipts for 1984 corporations with assets under $100,000 ($357.1 billion); and
- total net income (less deficit) for 1984 mining corporations with total assets $50,000,000 or more ($779.4 million).

Originally, such data were obtained by selecting all corporate returns with certainty. Then as the size of the population grew more unwieldy, sampling was introduced. At first, this was done by hand at about 64 district offices using many persons unskilled in statistics. Naturally, the design had to be kept as simple as possible. As computer assistance became available and the number of sampling locations decreased to the present ten Internal Revenue Service Centers, it was possible to add more stratifying variables and make the design more complex.

The population of corporation returns is highly skewed, with a relatively few large corporations accounting for well over half of the assets and income. In 1984, for example, the smaller 56% of corporations accounted for only 0.5% of the U.S. total assets, while the top 0.11% of the corporations accounted for 75% of the U.S. total assets. The sample design is therefore stratified, and the very large corporations are selected with certainty.

A fairly thorough, though highly condensed, documentation of the present sample design is provided in **Statistics of Income — 1984 Corporation Income Tax Returns.** Further descriptions of the design and estimating methods can be found in Harte (1982), Jones and McMahon (1984), and Oh and Scheuren (1987).

Under the current (1985) design, two stratifying variables are employed: Total Assets (TA) and Net Income or Deficit (NI). The former provides a measure of the level of total assets and other balance sheet items on the tax return and the latter is used to measure the size of income statement items which make up total income and total deductions. Each stratifying variable has a distinct set of classes and the strata are defined by the larger of the two classes. The type of business is also used in strata definition, especially in the larger size categories, because of the great number of financial industries having very high total assets, as compared to those of nonfinancial industries. The sampling rates are, then, determined using Neyman allocation. We treat the sample as if it were a stratified random sample but, in fact, it can be thought of as a stratified Poisson sample (as defined by Sunter, 1986).

3. ESTIMATING YEAR-TO-YEAR CHANGE

The features of the design described, so far, are concerned with the primary objective: to obtain accurate annual estimates; however, with samples being taken every year, a reasonable secondary objective is to estimate change over the years. Some provision for this has been made in the sample design, and this section will look at how change estimation is done and some innovations to improve these estimates.

For illustrative purposes, let us take the "change in total assets from one year to the next" as our variable of interest for measuring change and denote it by ΔTA. We simplify the discussion by considering only that part of the population that is subject to sampling (i.e., smaller corporations), and by considering, in this section, only TA as a stratifying variable. Table 1 shows the TA strata and their sampling rates for 1984 and 1985.

Table 1

Sampling Rates Based on Total Assets (TA)

Strata	TA (Total Assets in ,000)	Sampling Rates	
		1984	1985
(1)	(2)	(3)	(4)
1	0 - $50,	.002	.002
2	$50, - $100,	.002	.002
3	$100, - $250,	.007	.006
4	$250, - $500,	.01	.01
5	$500, - $1,000,	.02	.02
6	$1,000, - $2,500,	.07	.06
7	$2,500, - $5,000,	.1	.1
8	$5,000, - $10,000,	.2	.2
9	$10,000, - $25,000,	.7	.6

For estimating ΔTA, we are interested in classes defined by cross-tabulating across both years (Table 2). The first row indicates births, or new corporations in 1985. The first column indicates deaths, corporations no longer in business in 1985, or corporations that have merged. The center of the table shows the change classes for corporations in the population in both years. The cells on the diagonal represent corporations that were in the same stratum from one year to the next. Cells off the diagonal represent corporations that changed strata between 1984 and 1985.

If we were designing the 1985 sample to estimate ΔTA, we would want (1) a representative sample of births, and (2) as much sample overlap as possible for corporations existing in both 1984 and 1985. Ideally, to measure the overlap population, we would like to select the same corporations in both years' samples. If the entire 1984 sample of continuing corporations could not be used, we would probably, at least, want to emphasize sampling off the diagonal (i.e., selecting corporations that have large changes).

The cross-sectional design results in representative sampling of births; however, if left to chance, for corporations existing in both years, there would be very little overlap in the sample from year to year, except for large, static corporations which were taken at a 100% rate. If drawn independently, the effective sampling rate for selecting a corporation into the sample in both years is the product of the two years' sampling rates. Take, as an example, the cell representing corporations in stratum 1 in 1984 (TA under $50,000) and in stratum 3 in 1985 (TA from $100,000 under $250,000). The effective rate for selecting such corporations in both years' samples would be .000012 = (.002)*(.006).

Table 2.

Total Assets (TA) Cross-Tabulated for 1984 and 1985 Strata

		1985 Strata									1984 Sampling Rates
	Deaths	1	2	3	4	5	6	7	8	9	
	Births										n.a.
1984 Strata	1	....									.002
	2		....								.002
	3			....							.007
	4				....						.01
	5					....					.02
	6						....				.07
	7							....			.1
	8								....		.2
	9									....	.7
	1985 Rates	n.a	.002	.002	.006	.01	.02	.06	.1	.2	.6	n.a.

Note: n.a. = not applicable

The corporate sample design addresses the objective of estimating change by assuring a much larger overlap from year to year. Each corporation has an employer identification number (EIN). A pseudorandom number is associated permanently with each EIN. The corporation is selected for the sample if the random number is less than a constant controlling the sample rate. Therefore, if the corporation is selected on one occasion it will be selected again if the sample rate is at least as high. (This type of procedure is discussed in Harte, 1986). Using the EIN, the effective selection rate for a corporation being in both samples is the minimum of the two years' sampling rates. Continuing the previous example, the effective sampling rate using the EIN would be the minimum of (.002, .006) = .002, compared to the minuscule sampling rate of .000012 if left to chance.

Sunter (1986) calls such a procedure implicit longitudinal sampling and shows that it maximizes the overlap of units sampled on two or more occasions; but when most changes are small it still results in most of the sample overlap being on the diagonal. As Table 3 shows, emphasis is not placed on corporations with great change.

Cells above the diagonal represent corporations that grew from 1984 to 1985. The sample overlap here is small because the sampling rates in 1984 were smaller. To increase overlap we would need to predict in 1984 which corporations would grow in the future. It is doubtful whether we will ever be able to do this effectively.

Cells below the diagonal represent corporations with smaller TA in 1985, so the 1985 selection rate is smaller than the 1984. Therefore, many of these corporations would be in the 1984 sample but not in the 1985 sample. We can improve the overlap here by looking back to 1984 results before sampling in 1985.

In the last several years, stratifying variables have been added to the design to increase the number of corporations in both samples by "looking back" in this way. For example, a recent design change was to use the maximum of Total Assets and Beginning Assets as the stratifying variable, instead of just Total Assets. This "looks back" to 1984 -- because the 1985 Beginning Assets should equal the 1984 Total Assets --and would, therefore, increase the sample overlap below the diagonal.

Table 3

Effective Rates for Sample Overlap using the Employer Identification Number (EIN) as Seed

1984 Strata	1985 Strata									1984 Sampling Rates
	1	2	3	4	5	6	7	8	9	
1	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002
2	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002
3	.002	.002	.006	.007	.007	.007	.007	.007	.007	.007
4	.002	.002	006	.01	01	.01	.01	.01	.01	.01
5	.002	.002	.006	.01	.02	.02	02	.02	.02	.02
6	.002	.002	.006	.01	.02	.06	.07	.07	.07	.07
7	.002	.002	.006	.01	.02	.06	.1	.1	.1	.1
8	.002	.002	.006	.01	.02	.06	.1	.2	.2	.2
9	.002	.002	.006	.01	.02	.06	.1	.2	.6	.7
1985 Rates	.002	.002	.006	.01	.02	.06	.1	.2	.6	n.a.

Notes: n.a. = not applicable
Diagonal rates are underlined

4. DESIGN EFFECTS

As already discussed, modifications have been made to improve estimates of year-to-year change and other revisions are being considered for the future. In looking at these various options (both practical and theoretical) for changing the design, a necessary concern is what would be the design effect on the cross-sectional (annual) estimates? What would be the increase in variance for the annual estimates of TA and NI?

In this section, we calculate the design effects in the classical setting of stratifying on several variables, as described in Cochran (1977). Three variables are considered: TA, NI

and ΔTA (the change in TA from 1984 to 1985). Recall that the first two are the stratifying variables for the current cross-sectional design.

It was felt that using Net Income/Deficit (NI) as a stratifying variable may already control for ΔTA, because there should be a strong correlation between NI and ΔTA. The strength of this correlation is still an area for investigation. For this calculation, we did assume a modest ($r \approx 0.5$) correlation and, therefore, the results may be more favorable to the current design than is realistic. We need to look at these calculations under other models, as well.

In our analysis, we used the same technique for determining sampling rates as is used for the corporate sample. The standard deviation within each stratum is assumed to be proportional to the range of values in that stratum, and Neyman allocation is employed. Whenever TA and NI are used together, the stratifying variable is, in fact, the larger of the two classes; otherwise, when using more than one variable for stratification, the sampling rates derived from the standard deviations are averaged.

We looked at the seven possible designs, using combinations of these three items as stratifying variables. Figure 1 illustrates these possible designs. The vertices represent the sample designs based on the single variables; the bottom vertex corresponds to the sample design based on ΔTA only, etc. The points between vertices correspond to designs based on combinations of these variables. The designs of interest are:

- the design based on all 3 variables (TA, NI, and ΔTA), which represents the design options under current consideration;
- the design based on TA and NI, which represents the current design; and
- the three designs based on the single variables which represent the optimal design for each variable.

Using this diagram, we look at the design effects on the three variables, i.e., the relative variances under each design for estimating TA, NI, and ΔTA. (The variances calculated are conditional variances given the achieved strata sample sizes.) The designs are compared to the design under consideration -- the design using all 3 stratifying variables; so, by definition, this design always has design effect 1.00.

Figure 2 shows the design effects (deffs) for estimating TA. The "optimal" design for estimating TA is the design stratifying on TA alone, so it has the smallest design effect (0.96). The design effect of using only TA as the stratifying variable is the ratio of the variance of the estimated TA using the TA design, to the variance of the estimated TA using the design stratified on TA, NI, and ΔTA. The current design (based on TA and NI) is not optimal but results in an increase in the variance of TA (deff = 1.04). But note that this is greater than the design effect for the design using all three stratifying variables (1.00). So, adding ΔTA to the design could improve the estimation of TA over the current design.

Figure 3 shows the design effects for estimating NI. Adding ΔTA to the current design appears to have little effect on the variance of the estimated NI; the design effects are essentially the same.

Figure 4 shows the design effects for estimating ΔTA. The optimal design here would be to stratify on ΔTA only, and this design has the smallest design effect. Note how well the current design does for estimating change: only an 11% increase in variance over the optimal. (This assumes, of course, that we could look at both years and achieve complete sample overlap.) This is highly dependent on our assumption of moderate correlation between NI and ΔTA; if this correlation is overstated, then this deff is underestimated, and the variance of the estimated ΔTA may be much larger. The design using all 3 variables for stratification, which has a 3.5% increase in variance over the optimal, improves on the current design.

These results indicate that we can consider options to improve estimates of ΔTA (year-to-year change) without necessarily jeopardizing the annual, cross-sectional estimates. In fact, in all three cases here, the design under consideration was about as good as, or better than the current design.

Figure 1. - - Design Options

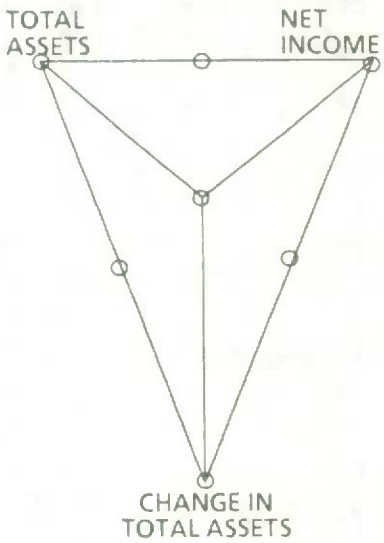


Figure 2. - - The Design Effects for Estimating TA

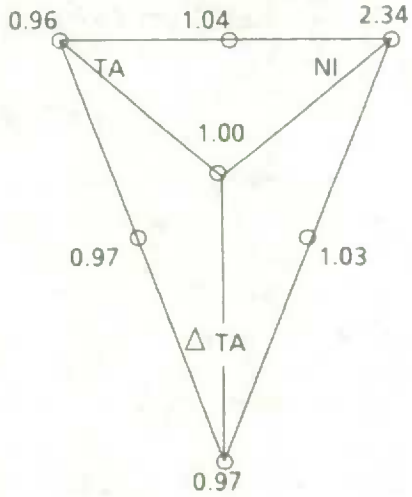


Figure 3. - - The Design Effects for Estimating NI

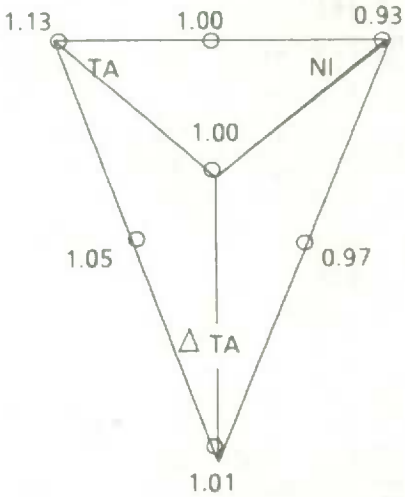
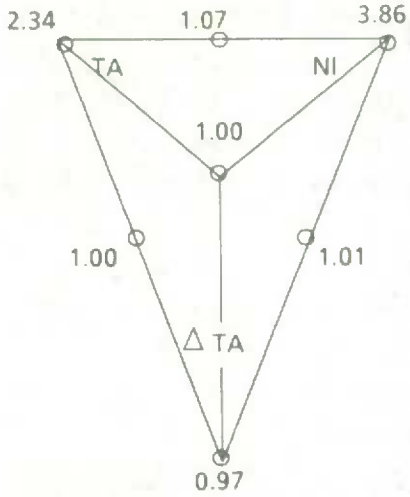


Figure 4. - - The Design Effects for Estimating Δ TA



5. CONCLUSIONS AND AREAS FOR FURTHER STUDY

In the above discussion we have been looking only at a static population, in the sense that we have not considered the effects of inflation or real growth. The strata boundaries are essentially fixed, relatively static through the years. This is done, partially, because our estimates are published every year for the population classes defined by these strata.

However, 3% inflation, for example, would cause movement across these strata boundaries that is not indicative of a real change in a corporation. Therefore, another option being considered is adjusting the strata cut points so that strata more nearly represent the same part of the population from year to year and movement out of a stratum is more indicative of real change. Such a modification could also improve the overall annual estimates; what might suffer would be the published estimates of the specific fixed dollar subclasses.

This has been a very brief overview of general ideas currently of interest for improving the corporate sample design; we have not discussed many of the real life difficulties of application, such as effects of nonsampling errors, controlling sample sizes, etc. Modifications have been made to the corporate sample design to improve the estimation of year-to-year change and we feel that there are other improvements that can be made, without jeopardizing the cross-sectional estimates. We are currently considering these modifications in more detail and hope to report on them further at the upcoming (1988) American Statistical Association meeting in New Orleans.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*, (3rd ed.) New York: John Wiley, 85.

Harte, J.M. (1982). Post-Stratification Approaches in the Corporation Statistics of Income Program, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 250-253.

Harte, J.M. (1986). Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 603-608.

Internal Revenue Service (1987). Description of the Sample and Limitations of the Data, *Statistics of Income - 1984, Corporation Income Tax Returns*, Publ. 16, Washington, D.C., 7-14.

Jones, H.W., and McMahon, P. (1984). Sampling Corporation Income Tax Returns for Statistics of Income, 1951 to Present. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 437-442.

Oh, H.L., and Scheuren, F.J. (1987). Modified Raking Ratio Estimation. *Survey Methodology Journal*, vol. 13, no. 2, Statistics Canada.

Sunter, A.B. (1986). Implicit Longitudinal Files: A Useful Technique, *Journal of Official Statistics*, vol. 2, no. 2, Statistics Sweden, 161-168. See especially 164.

Proceedings:
Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

THE USE OF ADMINISTRATIVE DATA FOR INITIAL AND SUBSEQUENT PROFILES OF ECONOMIC ENTITIES

COLLEEN CLARK and ROBERT LUSSIER[1]

ABSTRACT

Statistics Canada is currently rebuilding its central register of economic entities. The new register views each economic entity as a network of legal and operating entities which define statistical entities. This network view, the profile, is determined through the 'profiling' process which involves contact with the economic entity. In 1986 a list of all entities in-scope for a profiling contact was required so that profiles could be obtained to initialize the new register. Administrative data were used to build this list. In the future, administrative data will be a source of information on changes that may have happened to economic entities. They may thus be used to request review and updating of profiles.

The paper begins with the objectives of the profiling process. The procedures for constructing the frame for the initial profiling process using several administrative data sources are then presented. These procedures include the application of concepts, the detection of overlap between sources, and the evaluation of data quality. Next, the role of administrative data in providing information on changes to business entities and in requesting profiles to be verified is presented. Then the results of a simulation study done to assess this role are reviewed. Finally, the paper concludes with a series of questions on the methodology of using administrative data to maintain profiles.

1. INTRODUCTION

Statistics Canada is in the process of reorganizing its programme of economic surveys. The new programme will result in an increased use of administrative data. They will be part of a Central Frame Data Base (CFDB) from which economic surveys will draw their sample.

They will also be used to maintain the CFDB. This and other elements of the reorganization strategy are contained in Colledge and Lussier (1985). Experiences in the implementation of the strategy are contained in Colledge (1987).

[1] Colleen Clark, Social Survey Methods Division, Statistics Canada, 4-C1 Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6; and Robert Lussier, Business Survey Methods Division, Statistics Canada, 11-M R.H.Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

One of the first steps in the reorganization was to formulate definitions of the CFDB units. A fundamental unit is the business entity. A business entity is defined in Statistics Canada (1987) as 'an economic transactor having the responsibility and authority to allocate resources in the production of goods and/or services, thereby directing and managing the receipt and disposition of income, the accumulation of property, the borrowing and lending of capital, and the maintenance of complete financial statements accounting for their responsibilities'.

The reader should note that the term 'business', in the United States is often restricted to the trade sector. This is not the case in Canada. Business entities cover all sectors of the economy including areas such as manufacturing, transportation, and professionals. The term 'economic' was used in the title to avoid a potential misunderstanding. However the term 'business' is used throughout the paper.

The Central Frame Data Base currently being built by Statistics Canada attempts to represent the structure of the Canadian economy. It recognizes that this economy is dominated by a small number of large business entities who account for the majority of the activity within the economy. The CFDB is divided into two components paralleling this dichotomy.

One component, the Integrated Portion (IP), provides coverage of the small number of large or otherwise important business entities, while the other, the Non-Integrated Portion (NIP), covers the remaining large number of smaller entities. The entities in the former component are more complex. Hence, the identification of those portions of the complex business entity that are of interest to a particular survey requires substantial effort.

The Integrated Portion (IP) of the CFDB attempts to represent the complex structure of business entities through the use of an Information Model. The model consists of five structures linked together which describe a business entity. These structures allow survey populations to be accurately identified. Entities on three of the structures are not controlled by Statistics Canada while entities on the other two are generated by Statistics Canada for the purpose of collecting, editing, estimating, and tabulating economic data. The five structures are:

i. The **legal structure** which describes the legal representation of the business entity. It is comprised of legal entities and their relationships of ownership and control. Examples of legal entities are incorporations under federal or provincial charter.

ii. The **operating structure** which describes how the business entity operates and how it organizes its accounting system. It is comprised of operating entities. This structure organizes and controls the production of goods and/or services. It is an attempt to structure the business entity as it sees itself. Examples of operating entities are divisions, profit centres, and plants.

iii. The **statistical structure** which consists of a hierarchy of statistical entities. These entities are derived from the associated operating structure depending on the units within the operating structure for which records for a particular set of data are maintained.

iv. The **reporting structure** which consists of reporting arrangements for each selected statistical entity by survey. The data available in the accounting system of the business entity are collected from the reporting entities.

v. The **administrative structure** which contains administrative data such as income tax data collected from legal entities and payroll deduction account data collected from operating entities.

The complex process of determining the boundaries of the business entity and of delineating its five IP structures and their associated links is termed 'profiling'. This network view of the business entity is the 'profile'. The data to construct a profile are obtained through a contact with the business entity or some component of it. The entity's legal and operating structures as well as some administrative structure data items are obtained, or, reviewed and updated during the interview. The statistical structure is then generated or updated automatically from the new operating structure. Finally, default reporting entities are created for new selected statistical entities using selected fields from the legal, operating or administrative structures. These entities may subsequently be updated as a result of the first survey contact with the respondents.

The type of profiling contact used depends on the entity's complexity and any special reporting arrangements. The most complex and important entities will receive a personal visit from either Head Office or Regional Office personnel. The remaining entities will be contacted by telephone. Entities will be contacted about once every two years, or more often, depending on how quickly their structures change.

Cyclical profiling, whereby business entities are periodically contacted, is one method that will be used to keep the IP of the CFDB current. A survey feedback process and data from administrative sources will also be used.

The design and construction of the CFDB is taking place over three years culminating in a data base that will be available for integration into survey programs in April 1988. Most of the data in the Integrated Portion of the CFDB in April 1988 will have come from a profiling process that began in April 1986. However, no single list of business entities in-scope for a profile was available in April 1986.

Administrative data played a major role in initiating the profiling process. It was used as a starting point to construct the current Statistics Canada view of the business entity. A list of business entities in-scope for an initial profile was assembled from administrative data sources. Section 2 describes how this was accomplished. Section 2.1 gives the frame requirements. A description of the data sources used to build the frame follows in Section 2.2. Section 2.3 shows how the frame unit was constructed and how the various data sources were combined to build the frame.

Section 3 describes how administrative data will be used to detect potential changes in a business entity and then to initiate the maintenance profiling process. The results of a simulation study done to quantify the proposed use of administrative data sources are then presented. The paper concludes with a discussion of several issues that this study has raised.

2. USE OF ADMINISTRATIVE DATA FOR INITIAL PROFILING

2.1 Frame Requirements

The first step in building the frame for initial profiling was to define the frame unit. The ideal one would be the business entity. However this entity was not available either internally or externally to Statistics Canada. The units available to us were essentially legal entities. It was necessary, then, to group legal entities to approximate business entities. The frame unit was defined as a grouping of legal entities subject to the following constraints:

i. The definition of the business entity implies that it covers all legal entities linked through control. One type of control is established through owning more than 50% of the voting rights of a legal entity. The grouping of legal entities through this control rule is restricted to one level of foreign control outside Canada.

ii. There has to be a single Canadian legal entity that controls all other Canadian legal entities in the business entity. This is necessary because profiling contacts with the business entity could only be made in Canada.

The next step was to determine which frame units would comprise the frame and what data was required for each. The frame from which business entities would be selected for an initial profiling contact and from which the initial picture of the business entity would be generated would contain all business entities in-scope for a contact.

Business entities are in-scope for a profiling contact if they qualify to be members of the Integrated Portion of the CFDB. Membership is determined by criteria applied to the legal structure that describes the legal representation of the business entity.

Legal structures can become members of the Integrated Portion in one of two ways. First, if the structure consists of only one legal entity then the legal entity is part of the Integrated Portion if its revenue during its fiscal year of interest is above a set prespecified value. This prespecified value depends on the legal entity's major industry and the location of its head office. Alternatively, if the legal structure consists of more than one legal entity then the legal structure is part of the Integrated Portion if at least one of the legal entities in the structure is above its appropriate prespecified value.

Therefore, in order to determine which business entities are in-scope, the following information was required for every legal entity:

i. Relationships of ownership between legal entities.

ii. Revenue in the fiscal year of interest, primary industry, and head office location.

For business entities that qualify to be on the frame and, hence, to receive an initial profiling contact, information was required to select and contact the entity. The following was required to select the entity:

i. All industries in which the business entity was involved so that the Wholesale and/or Retail industries could be contacted first. The surveys of these industries required a set of statistical entities that had been generated from a profiling contact before other surveys did.

ii. The number of physical locations of all business entities that consist of one legal entity or that consist of two legal entities of which the owner is foreign. This data item determined the type of profiling contact that would be made as either a telephone contact by Regional Office staff or a personal visit by Regional or Head Office staff.

iii. The province in which the ultimate Canadian owner was located. The province was used to distribute the workload of making the profiling contacts to regional offices according to their capacities.

In order to contact the business entities, name and address were required for the legal entity at the top (excluding foreign owners) of the business entity. Contact data and any special reporting arrangements that surveys had recently used would be desirable.

2.2 Data Sources

The data sources which could be used were restricted, primarily, by the frame coverage requirements. This restriction eliminated sample lists and many industry specific lists such as survey frames. Only data sources that were lists of all legal entities potentially in-scope for a profiling contact that carried, at least, some of the required data items could be considered. These data sources were:

i. The **Inter-Corporate Ownership Database** (ICO) which is a list of all legal entities operating in Canada that are owned by either foreign or Canadian legal entities and their owners. The coverage of foreign legal entities is to the extent required to determine the ultimate owner.

ii. The **Current Business Register** (BR) which is primarily a list of all legal entities that are employers. The number of physical locations of a legal entity, contact data (address and reporting arrangements) used by surveys, and the industries in which the legal entity operates are available here.

iii. The **Corporation Tax Base** (CORP) which is a list of all legal entities that filed a corporate tax return with Revenue Canada, Taxation in a given year. The primary industry, the location of the Head Office, and revenue for the fiscal year are carried on this data source.

iv. The **Individual Tax Base** (IND) which is a list of all individuals who filed a tax return with Revenue Canada, Taxation in a given year. Individuals who report self-employed income on their return are legal entities of interest to Statistics Canada economic surveys. Primary industry data and contact data are available from this tax base for each individual reporting self-employed revenue as is his/her revenue from self-employment.

Both of the tax base data sources (CORP and IND) are administrative data files. Administrative data received monthly from Revenue Canada, Taxation regarding an employer's payroll deductions are used to update the BR. The ICO data source is a census survey response file.

None of these data sources provides complete coverage and all the required data items. Rather, coverage can only be obtained by combining these data sources. The same is true for some required data items while for the rest more than one source can provide them. The strategy used to combine these data sources to obtain the best coverage and data quality is presented in the next section.

2.3 Frame Creation Procedures

The challenge in creating the frame for initial profiling contacts lay in integrating four data sources that had each been designed for different purposes and had never been integrated to this extent before. This situation is common to users of administrative data. The task was even more complex because this was the first time many concepts established for the CFDB were applied.

The constraints of limited time and resources forced the project team to make some assumptions when creating the frame. However, the assumptions were justifiable since the picture used on the frame would be corrected through the profiling process. A simple description of the procedures used is presented in this section.

There were three steps in the frame creation process, each of which is discussed in the following sections.

i. Construct a list of all potential frame units;

ii. Determine which are in-scope; and

iii. Acquire selection and contact data.

2.3.1 Create Potential Frame Units

The frame unit was constructed by grouping legal entities in the following manner to create business entities. The legal entities were first grouped into legal structures. One

legal structure consisted of that set of legal entities related via ownership of more than 50%. Relationships involving foreign legal entities were accepted only if the foreign legal entity owned or was owned by a Canadian legal entity. When a foreign entity owned more than one Canadian entity, the legal structure was divided into as many business entities as there were Canadian entities directly owned by the foreign entity. In this way, a profiling contact would be made with the ultimate Canadian owner of each resulting business entity. Examples are provided in Diagram 1 at the end of Section 2.3.1.

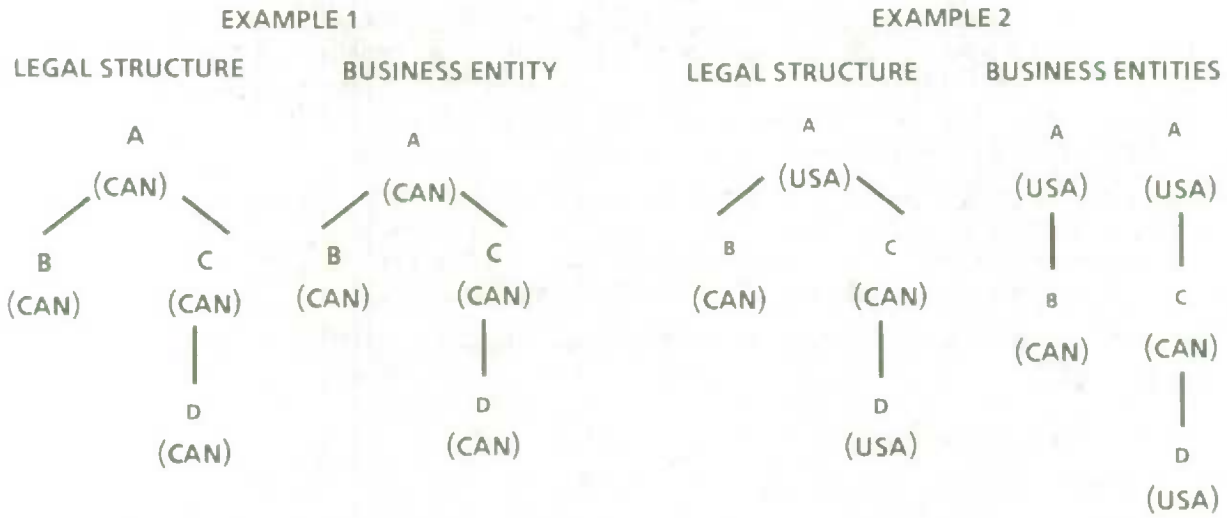
Individuals who reported self-employed income were considered as a legal structure containing only one legal entity. The ownership of corporations by individuals as well as relationships of joint venture between corporations were not considered in constructing business entities.

Therefore, we can think of the set of business entities in-scope for an initial profiling contact as two mutually exclusive groups. The first group consists of legal entities that represent individuals who report self-employed income. The Individual (IND) tax base contains a list of all potential frame units in this group.

The second group consists of legal entities that represent corporations operating in Canada. The Inter-Corporate Ownership (ICO) data source was manipulated to provide a list of corporations that belonged to legal structures containing more than one legal entity. A list of all legal entities that are not owned by any other legal entity was obtained from the Corporation tax base after elimination of those legal entities that were owned by other legal entities or were owners themselves. That is, it was necessary to match the ICO source and the CORP Tax base to identify the overlap between them. Legal entities that appeared on both sources could thus be identified to ensure that they would only appear once on the frame. Linkage between the two sources was not straightforward and involved a clerical process because a common identification number was often not available.

DIAGRAM 1

DEFINING BUSINESS ENTITIES



2.3.2 Determine In-Scope Frame Units

The data required to determine if individuals reporting self-employed income were in-scope was on the IND tax base. It was a simple step to determine if a legal entity was above its appropriate prespecified cut-off.

The situation was more complex for corporations. The linkage achieved between ICO and CORP provided the data required to apply the cut-off rule. However, about 20% of the corporations on ICO could not be linked to CORP. In these cases an assumption was made which led to an overestimation of the set of business entities in-scope for an initial profile. It was assumed that legal structures which contained at least one unlinked corporation satisfied the frame inclusion conditions. Otherwise, legal structures were frame members if at least one corporation satisfied the cut-off rule.

2.3.3 Acquire Selection and Contact Data

The result of the previous step was a proxy list of all business entities in-scope for an initial profiling contact. The data required for selection and contact described in Section 2.1 that are not already on the frame were available from the BR. The frame and the BR overlap because a majority of the frame units representing corporations and a smaller proportion of the frame units representing individuals are employers. Linkage between the frame and the BR was required so that data from the BR could be added to the frame for units found on both sources. That is, it was necessary to detect duplication between the two sources.

It was even more difficult to link these two sources than it had been to link the ICO and CORP sources. This was due not only to the frequent absence of common identification numbers as in the ICO-CORP case but also because the BR resembles a business entity's operating structure more than its legal structure. The name and address from the BR were used for linking when no common identification number was available. However, the names and addresses on the BR often refer to 'trade' or 'operating' locations which are sometimes different from the 'legal' names and addresses on the ICO and CORP sources. When this occurred it was difficult to establish a link and hence eliminate duplication.

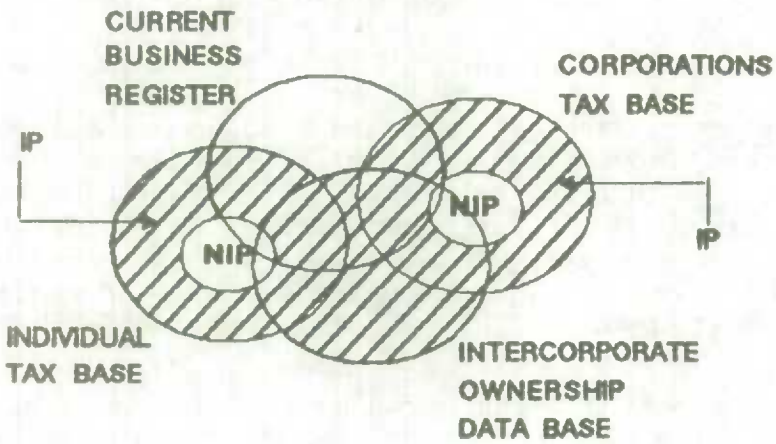
There were some frame units for which no link to the BR was achieved either because they were non-employers and therefore not on the BR or the linkage procedures could not establish the link. In these cases subsequent stages in the initial profiling process were amended to accommodate the frame limitations. Contact data of a lesser quality were taken from the tax base. The selection criteria were changed to reflect the absence of data on industrial breakdown and physical locations for these legal entities.

When a legal entity was involved in only one industry, the primary industry was available from both the tax bases and the BR. It was necessary, then, to reconcile this common data item when they were different. In this case the BR industry was used since it was considered more reliable.

A pictorial representation (not to scale) of the resulting frame is shown in Diagram 2.

DIAGRAM 2

FRAME FOR INITIAL PROFILING



2.3.4 Evaluate the Frame

The quality of the resulting frame was assessed by three projects. First, the consistency of the frame with the specifications for creating it was verified.

The second project involved comparing various distributions of the legal entities on the frame with the same distributions produced from an independent simulation of the Integrated Portion. The distributions did not differ significantly.

Lastly the frame was assessed by comparing it with the BR. A sample of 30 of the larger units in the BR was matched to the frame for initial profiling. All of the entities were found but with great difficulty because the two sources use different concepts.

2.4 Conclusion

The frame strategy just described was based on some simplistic assumptions regarding coverage, data quality, and the way in which business entities operate. 'Shortcuts' were often used to satisfy the frame requirements. It was felt that this approach was justified because of the role of the frame as a provider of initial pictures of business entities that would be updated during the profiling process. The implications of making these assumptions are discussed in this section.

The population of business entities in-scope for an initial profiling contact may contain duplicates and out-of-scope units. If so, then more profiling contacts than necessary will be made. This would increase Statistics Canada's production costs. It would unduly burden the respondent with duplicate requests. Finally, the image of Statistics Canada could be adversely affected.

The population may be underestimated. Nevertheless, the missing units will be profiled at a later date. This would delay the introduction of new large units into the Integrated Portion of the CFDB. The missing units would be covered by the Non-Integrated Portion in the interim rather than the Integrated Portion.

Inaccurate selection and/or contact data could complicate or delay contact until accurate data could be found. The consequence in these cases is also an inaccurate CFDB until the profile is completed.

These experiences demonstrate the complications introduced when administrative data are used. They also illustrate the care that must be taken in ensuring the compatibility of administrative data with one's requirements. Examples were provided of the types of ensuing compromises that must be made when a reasonable compatibility cannot be reached.

3. USE OF ADMINISTRATIVE DATA IN SUBSEQUENT MAINTENANCE PROFILES

3.1 Cyclical and Reaction Profiling

There will be two types of subsequent maintenance profiling, namely cyclical and reaction profiling. Each of these is explained below.

Cyclical profiling is the process that will ensure that all business entities in the profile population get reprofiled within a certain period of time. It is expected according to current budget forecasts that this period of time will be two years. Time elapsed since the business entity's last profile will be the factor that determines eligibility for cyclical profiling. Other factors will be taken into account to prioritize the eligible units within cyclical profiling.

Reaction profiling is the process that will profile a business entity as a result of information through a source other than profiling that changes may have occurred to that business entity and that the statistical image of the business entity on the register may not be valid any longer. Reaction profiling will keep the CFDB more up-to-date than if only the cyclical profiling mechanism were used. Some of the sources of information on changes are the various files of administrative data received regularly at Statistics Canada.

3.2 Sources of Administrative Data that can be used

The three sources of administrative data that Statistics Canada can use to update its central register that are discussed in this paper are:

- the Individual Tax Base;
- the Corporation Tax Base; and
- data on payroll deduction accounts captured by the tax authorities.

Generally, individuals and corporations file a single tax return for a reference year. However, it is possible to have more than one return for a reference year if, for example, a corporation changed its fiscal year end with the approval of the tax authorities. Nevertheless, one can say that tax returns are an annual source of changes.

The receipt of the tax bases at Statistics Canada does not occur at a single point in time. In fact, Statistics Canada receives files of tax data regularly for a reference year over a period of two years. Thus, one could perform monthly updates to the register from tax data but each register record would generally be updated only once a year.

On the other hand, an employer is generally expected to send remittances for his payroll deduction accounts on a monthly basis. In turn, Statistics Canada receives a file of payroll deduction account data once a month. Thus, monthly updates can be made to the register from payroll deduction account data and each register record can in theory be modified every month.

Note that there are other sources of administrative data that could be used. They are not discussed in this paper because they are not obtained on a universe basis or on a regular basis. They are nevertheless worth mentioning. These are:

- limited information on corporations that have not filed a tax return but are believed to be active, captured by the tax authorities;
- additional data captured from a sample of tax returns by Statistics Canada; and
- data on a tax authority form filled out by employers when they request a payroll deduction account, captured by Statistics Canada.

3.3 Signals of Change

Signals of change were developed from the administrative sources described in the previous section. These signals identify administrative records for which changes to their associated statistical entities may have occurred. They also inform the register that reaction profiling may be desirable for these entities to keep the register up-to-date.

The signals are administrative source dependent. For each of the three sources listed in 3.2 the signals consist of comparison tests between new data received for an administrative record and the last data received for the same record from the same source. These tests may involve a single field or a group of fields and may be conditional on a single field or several fields. These comparison tests attempt to identify real world events that have an impact on the statistical entities and not only on the administrative entities. Remember that the statistical entities exist for the purpose of economic statistical programs and often are completely different from the legal-administrative reality. Therefore, these comparison tests should optimize the detection of changes in the administrative data that reflect a change in the statistical entities. As an example, change of ownership of a manufacturing plant may mean the deathing of an administrative record and the birthing of a new one. On the statistical entities, it may however mean no change as the same establishment with its capabilities to provide the required data may still exist.

If the frame was updated directly from the changes noted in the administrative records, the consequence would be a high incidence of apparent deaths and births in the statistical entities and a risk of incomplete or duplicated coverage. Thus there is a requirement to contact respondents, or at least to perform in-house research using all available documentation, to find out for signaled administrative records what happened to the statistical entities. The "translation" process is not trivial at all and its resolution constitutes the purpose of reaction profiling.

The number of signals that were determined from each source together with some signal examples are presented in Table 1. One should however note the following points in studying the data on the number of signals. Some signals are very refined while others are not. It was often decided to split an original signal into mutually exclusive sub-signals because it was felt that it may be more informative in determining the action to take from the signal. The most trivial example concerns the Payroll Deduction Accounts. Eighteen of the 40 signals represent changes in the estimated number of employees covered by the account. The 18 signals distinguish between increases and decreases in the estimated number and the magnitude for each of them. It was thought that such a breakdown would be informative to prioritize the clerical work. Nevertheless one could consider these signals as one.

Table 1

Signals by Administrative Source

Administrative Source	Number Of Distinct Signals	Examples
Annual Individual Tax Returns	50	Change from single province of taxation to multiple jurisdiction
Annual Corporation Tax Returns	49	Start of a joint venture
Monthly Payroll Deduction Accounts	38	New account with descriptions in the name that identify a corporation

It is expected that even though tax returns are processed regularly, a given return will generally generate signals at most once per reference year while a given payroll deduction account may generate a signal or signals every month. What is of more interest therefore is not the number of signals defined per source but the number of records that are identified by these signals. This would give an idea of the amount of clerical resources that will have to be invested to update the register from administrative sources. A simulation study was thus undertaken to address this issue.

3.4 Simulation Study

The simulation study consisted of applying the signals previously described to the following populations:

- the individual tax returns for fiscal periods that ended in 1984 to detect changes that had taken place during these periods;
- the corporation tax returns for fiscal periods that ended in 1984 to detect changes that happened during these periods;
- the payroll deduction account of the beginning of October 1985 to detect changes that had occurred since the beginning of September 1985.

The results of the simulation study are presented in Table 2. The following observations can be made on the results:

- There are a very large number of tax returns that generate signals: only about one eighth of the individual tax returns and one fifth of the corporation tax returns do not generate any signals.
- There are 8,258 payroll deduction accounts that generated signals for a one month period. If one supposes uniformity of the payroll deduction account signals over months, there would be almost 100,000 accounts signaled in a year. Note that it is likely that accounts would be signaled in more than one month and therefore there would be duplicates if one cumulated the signals.
- If all records signaled in a year are added, it gives the grand total of 244,269 signaled records. However, it is obvious that signals are duplicated between the administrative sources. For example, a change to the legal name of a business could be found on the tax return as well as on each of its payroll deduction accounts.

Table 2

Results of Simulation Study

Administrative Source	Number In The Profile Population	Number Signaled	Percentage Signaled
Individual Tax Returns	72,190	63,446	87.9
Corporation Tax Returns	102,688	81,727	79.6
Payroll Deduction Accounts	134,973	8,258	6.1

3.5 Questions Raised

The results of the simulation study as well as an examination of the role of the signals raise a certain number of issues with respect to the profiling activities.

Six of these issues are presented below.

Performance of Signals in Detecting Change(s) to Statistical Entities

The signals will attempt to flag legal and/or operating entities involved in real world events that have an impact on the statistical entities. An update will then be necessary on the central register to maintain the quality of the statistical products. Are the signals really reflecting real world events that affect the statistical entities or are there some that have no impact? If some are useless, work will be generated for no purpose.

A small-scale survey was conducted in 1986 to determine the usefulness of the signals with respect to the detection of changes to the statistical entities. However, for various reasons, the only signals that could be used were those of the simulation study. They refer to changes between tax returns of taxation years 1983 and 1984. Thus the time lag between the reference period of the signals and the survey period (1986) gave recalling difficulties to the respondents. This led to the inclusion of events which took place after the period as well as the omission of events which did occur in the reference period. The survey was therefore inconclusive and no other attempt has been made since then.

Repetitiveness of Signals

Signals will be received over time and from different independent sources. The tax returns in particular suffer from noticeable time delays. As a given signal is received, the CFDB may have already been updated to reflect the real world event behind the signal. This update may have been the result of processing a signaled record from another source or of conducting cyclical profiling or of incorporating feedback received from surveys. Therefore, signals cannot be treated independently of the CFDB to decide to perform a reaction profile. However, how should a signal be checked against the CFDB to see if the CFDB was already updated? As an example, if a large increase in revenue is flagged on a corporation tax return, how should one check if the CFDB was already updated to reflect the real world event behind this increase when one does not know the real world event behind it?

Omission of Signals of Change

Similarly, some records will not get signaled. Will the absence of signals definitively mean that no real world event occurred that need the statistical structure to be updated? Should other signals be developed to cover omissions? Again, the survey previously mentioned was inconclusive in answering these questions.

Availability of Resources to Handle Signaled Records

As the simulation study showed, a large number of records will be signaled. These will require manual work. It is likely, that there will not be sufficient resources to perform all this work. How should the total amount of resources to be devoted to reaction profiles be determined and how should this total amount be used to handle the signaled records? If constraint on resources demand that some signals be ignored, how will these be determined?

Response Burden

The results of the simulation study suggest that businesses will be contacted more often than every second year to check for frame changes other than through regular survey activity. This will increase response burden. Can a trade-off be established between increase in response burden and out-datedness of the register? What should this trade-off be?

Role of Cyclical Profiling

The large amount of records signaled by the tax returns in the simulation study raises a question about the usefulness of cyclical profiling. The number of records subject to cyclical profiling and not to reaction profiling can be deduced to be very small. First, suppose the results of the simulation study in terms of numbers hold for a second year. Then suppose the records signaled in the second year are not all the same in the first year but that there are new records signaled and that there are last year's records not signaled the second year. Then it can be safely assumed that the number of records which will not get a signal over two years will be very small. There may be only a few records left which will not be signaled on either one or the other year. This will in fact represent the maximum target population for cyclical profiling. Will it be necessary to perform a profile for these entities, knowing that they are not signaled by the Payroll Deduction Accounts nor by the tax returns?

4. CONCLUSION

Section 2 has shown how administrative data were used to build a frame for initial profiling. Administrative data offered extensive coverage. However, it was also seen that conceptual differences between one's requirements and administrative data can lead to complications requiring simplifying assumptions and compromises.

The resulting frame supported the initial profiling of all business entities except the most complex ones. In these cases the approximation given by the frame could not be accepted. Rather, extensive research was conducted on each business entity using elements such as public annual reports and survey responses.

The frame also played an important role in initializing the CFDB. It was used along with the Business Register to identify the members of the Integrated Portion.

The method by which administrative data will be used to initiate a maintenance profile was described in Section 3. Signals of change will be derived from various administrative sources and will generate requests to verify profiles. Many issues were raised in this

respect. These issues are being addressed by the various design teams responsible for implementing the CFDB update strategy. A solution being investigated to solve some issues is to prioritize signals depending for example on the length of time since the entity was last profiled. Another solution is to develop a self-learning process. Experience will dictate which signals are useful and should be kept. Therefore, substantial work is still required before the process stabilizes in production.

REFERENCES

Colledge, M. and Lussier, R. (1985), "A Strategy for the Provision of Frame Data and Use of Tax Data for Economic Surveys", *Proceedings of the Section on Survey Methods*, 1985, American Statistical Association, Washington, D.C.

Colledge, M. (1987), "The Business Survey Redesign Project - Implementation of a New Strategy at Statistics Canada", presented at the Bureau of the Census Third Annual Research Conference.

Statistics Canada (1987), "Version 4.2 of the CFDB Data Dictionary", Business Survey Redesign Project Working Paper, March 4, 1987.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

INTEGRATING STUDENT RECORDS INTO A DYNAMIC DATABASE AND STATISTICAL REPORTING SYSTEM

ANN E. HOLLINGS and BRIAN D. PETTIGREW[1]

ABSTRACT

The Registrar's Office at the University of Guelph is a major source of data for research undertaken by the Student-Environment Study Group. These administrative data are structured to simplify the record-keeping role of the Registrar's Office by allowing easy access to individual student records. However, in this format, descriptive population information is difficult to extract, and it has been an obstacle to researchers and University planners that this wealth of data has been effectively inaccessible. A system is described which organizes these data into an integrated, cohesive database. This system can track student development from university entry to departure, support any combination of the Registrar's records, and simplify the selection of random stratified samples for surveys. By providing a link between survey results and the source data, it dispenses with the necessity of asking respondents to supply information already contained in the database. In many cases, the researchers' questions can be addressed without the need to conduct a survey at all, simply by analysing the appropriate subpopulation.

1. INTRODUCTION

High attrition rates, particularly in the first year of university, are of increasing concern to university administrators and policy makers. High attrition results in significant losses on a number of fronts, including financial loss to the institution, and personal and financial loss to the student (Gilbert and Gomme, 1986). At the University of Guelph, the Student-Environment Study Group (SESG) was formed to examine the learning environment of the undergraduate student population in an effort to identify and better understand potential problem areas, and to facilitate the process of policy formulation and program evaluation.

Typically, investigation of the student climate takes the form of surveying a relevant sample of the population in order to ascertain student attitudes, feelings and reactions to a number of issues. Selection of the relevant subpopulations and the presence of confounding factors in surveys of this type often necessitate collection of academic and personal data. Much of this information has already been compiled for university administrative purposes, and this was one of the primary motives for organizing the SESG Tracking System. This paper describes the process the SESG has established for using a

[1] A.E. Hollings and B.D.Pettigrew, Student-Environment Study Group, University of Guelph, Guelph, Ontario, Canada N1G 2W1

data base originally intended for record-keeping functions and constructing from it a cohesive and dynamic database which can track student progress from university entry to departure (by either graduation or withdrawal).

2. SOURCE OF DATA

Student records compiled for administrative purposes by the Registrar's office at the university, contribute most of the data stored on the tracking system. These records are originally structured in discrete parcels of information based on the kind of data provided; for example, all personal data such as name, birthdate and sex are provided on a personal record, while admission information such as Grade 13 average, high school attended and program selection, are located on the admission record. As a student progresses through university, some records are added (for instance, marks for each semester or applications to new programs) and others are updated (perhaps address or marital status). Retrospective or longitudinal studies of the student population are difficult using this data base directly, for although most of the information is retained in some form by the Registrar, it is often difficult to retrieve and unwieldy to use.

In response to various interdepartmental requests, the Registrar's office makes available a number of packages of information, containing selected record types for any semester. These options are presented in cross-sectional format, providing the required information for a given semester. This is extremely helpful and sufficient for most users of the data, but for the purposes of tracking student progress over the course of a university career, the difficulties in using an administrative database become clear. In its original form, the database is designed to deal efficiently with current records; for example, in what courses is a student currently registered, or what is a student's current address. Clearly, for tracking purposes, such a data structure presents some challenging logistical problems.

3. OBJECTIVES OF THE SESG TRACKING SYSTEM

The tracking system, a computer-based system of data files and computer programs, forms the foundation of resources from which the SESG regularly draws in order to meet its main objective of monitoring and better understanding all aspects of student life. The specific objectives of the tracking system can be summarized as follows:

1. To establish a database which accumulates personal and academic history on all students as they proceed through university;
2. To provide the capacity to build on the established database by including survey results on individual student records;
3. To track individual student progress through his or her university career;
4. To provide a means of assessing and describing population characteristics and trends; and,
5. To facilitate the drawing of random samples and stratified random samples from the population.

The Registrar's data, originally designed for the primary purpose of record-keeping, is the most significant resource used by the SESG in meeting these objectives. With continued co-operation between personnel from both the SESG and the Registrar's office, the tracking system has managed to maximize the potential of the resources available, while minimizing duplication of effort.

4. ADVANTAGES OF USING ADMINISTRATIVE DATA AS A BASE IN SURVEY WORK

In usual survey situations, historical and program information would be have to be solicited from the respondent directly in the survey. Some of the problems associated with this practice, and ways in which the tracking system can circumvent them include:

1. Data that rely on memory are often inaccurate (for instance, Grade 13 average). Sometimes even current information is inaccurate; for example, a student may declare his intended program of study rather than the program in which he is actually registered. In the tracking system the accurate records are already on file.
2. Questions designed to elicit personal information are sometimes regarded as an invasion of privacy by respondents, even though the information is used only in the production of statistics. The use of a control number protects the privacy of the individual but still permits the important link to the required information.
3. Duplication of effort is avoided by using data already collected by university administrative departments.
4. There is a reduced burden on the respondent to provide detailed information.
5. Possible confounding variables can be included with survey results; an unexpected factor may be discovered to have a significant influence on the results. Under normal conditions it would be impossible to request additional information from survey respondents.

Useful statistics can be generated without having to survey the population at all. For example, examination of first year course mark distributions, or tabulation of the number of changes of major areas of study, can provide enormously helpful insights to curriculum planners and department organizers. Such information can be compiled directly from the database.

5. WHAT IS THE TRACKING SYSTEM?

The tracking system is a collection of data files and computer programs designed to facilitate the statistical analysis and production of reports describing the student population at the university.

Figure 1 is a schematic representation of data sources and file structure in the tracking system. Among the data files are a number of master files, one created at the beginning of each semester. By the time the system has been in place for four years, there will be a total of 12 master files residing on the system (three a semester for four years), providing information on all students registered at the university. At that point, each semester when a new master file is created, the oldest master file will be retired to archival storage.

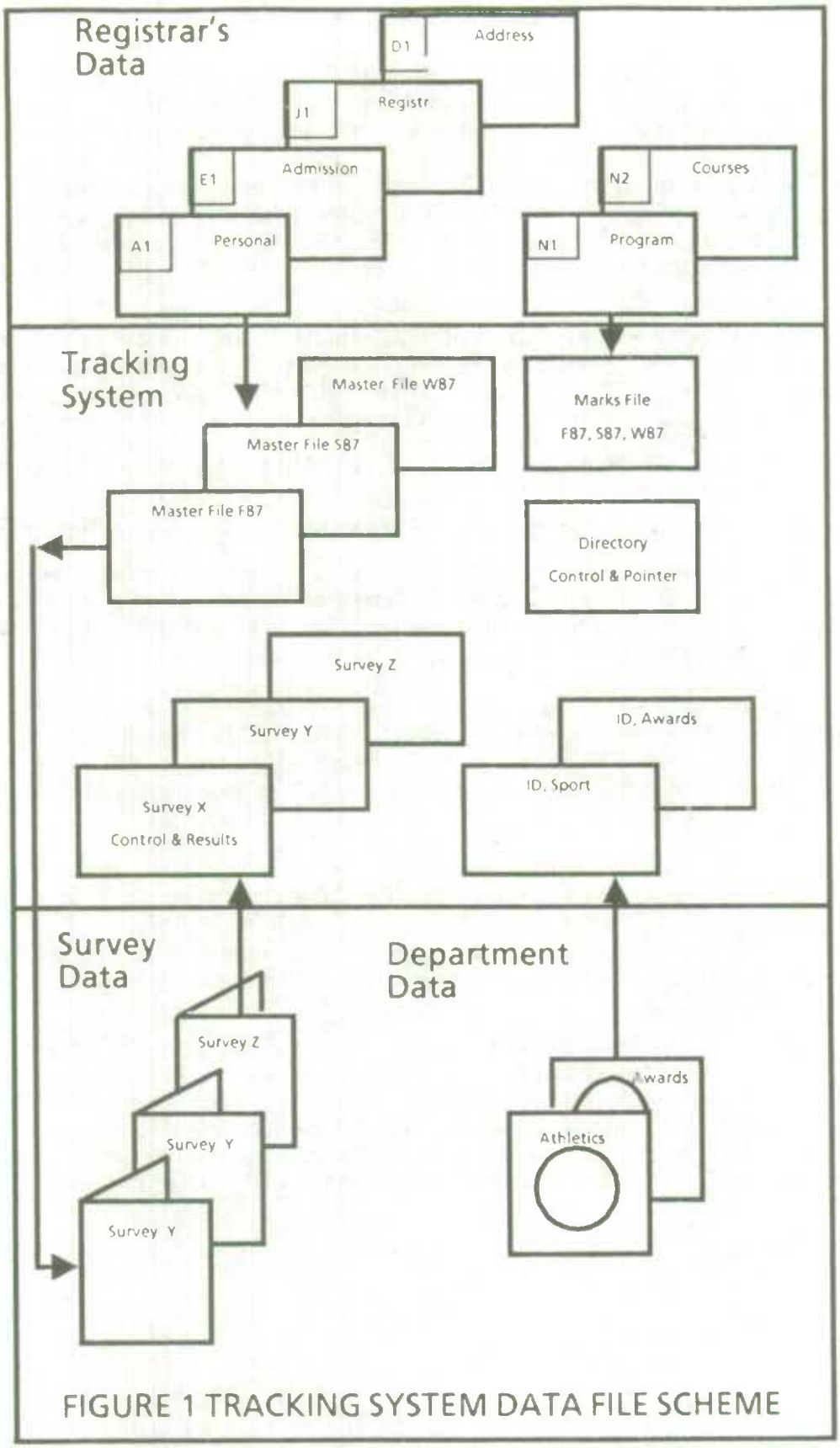


FIGURE 1 TRACKING SYSTEM DATA FILE SCHEME

At the beginning of each term, the SESG receives a computer tape from the Registrar containing names and admission data for all incoming students. Initially, all the original information is converted by the SESG from ASCII format to EBCDIC and stored on a tape in the SESG tape library. This way the information remains fairly accessible, if required for data rechecking, but is not using valuable computer space.

The information of particular interest to the SESG is collected and stored in a master file. A part of this process is the assigning of a unique control number to each new entry to the tracking system; the control number rather than the student identification number, provides the link between data files. The master file then, includes the control number and selected data from some of the Registrar's records. A number of blank fields designated as survey indicators are also built into the master file to provide references regarding particular surveys, specifically, who was sent Survey X and who responded to Survey X. It is then a very simple procedure to check who has been selected as part of a survey sample and to ensure that no-one is oversurveyed.

The master files are stored on the computer as MasterSYY, where S represents the semester and YY represents the year (for example, MasterF87 denotes the file containing data for students entering the university in the Fall of 1987). As soon as the marks for each semester become available, they are incorporated into a separate Marks file, which accumulates semester, average and program for all students in the tracking system. In addition there is a directory for reference purposes, giving control number and the master file containing each student's records.

Other data files on the system contain specific survey results, usually individual survey scores identified by means of a control number.

The immense utility of the tracking system lies in the unique free flow structure of the data files within the system itself. Any combination of master and survey files can be merged, so that the entire population or subsets of it with accompanying marks information, can be easily incorporated into a single data file. For example, the master files for Winter, Spring and Fall of 1987 could be merged to form a calender year cohort, or master files from Spring 1987, 1986 and 1985 could be combined for an examination of summer students. In both cases, the control numbers can be linked to the Marks file to provide a more complete picture of the population of interest.

6. RESOURCES

The SESG operates its tracking system using a combination of computer resources, ensuring a large degree of autonomy with respect to the university mainframe computer. The central machine is an IBM compatible PC with a 30 megabyte hard disk, loaded with a number of PC software packages including LOTUS, PC-SAS, WordPerfect and a terminal emulator. The tracking system data are stored on the mainframe computer, and are accessible on the PC through the terminal emulating software.

Many of the surveys, especially those dealing with small subsets of the population, can be easily analysed on the PC (using SAS), and the results can be stored on floppy disks. The 'bottom line' part of the survey can be uploaded to the mainframe and stored in a separate data file. PC-SAS is also useful for generating mailout labels and mailout lists, decreasing reliance on the mainframe utilities. A plotter links to both mainframe and PC software, allowing graphic representation of the data through a selection of mainframe and PC software.

The SESG mainframe workspace has a number of executive programs built into a menu-driven format, which considerably simplifies such operations as the conversion of the original data to SESG tape, the selection of subpopulations based on user-stipulated criteria, the merging of the master files with academic information, the random selection

of a sample, monitoring returns for follow-ups, and the updating of records with current address, marks and survey material.

7. EXAMPLE 1 — THE UNIVERSITY RESIDENCES ENVIRONMENT SCALE

One of the surveys used last winter by the SESG was the University Residences Environment Scale (Moos and Gerst, 1974), a survey designed to assess the social climate in residences on campus. The questionnaire itself is administered to individuals within the residences, and the scores are calculated for each residence hall. The procedure used to select the sample, enter the returns and analyse the results is shown schematically in Figure 2.

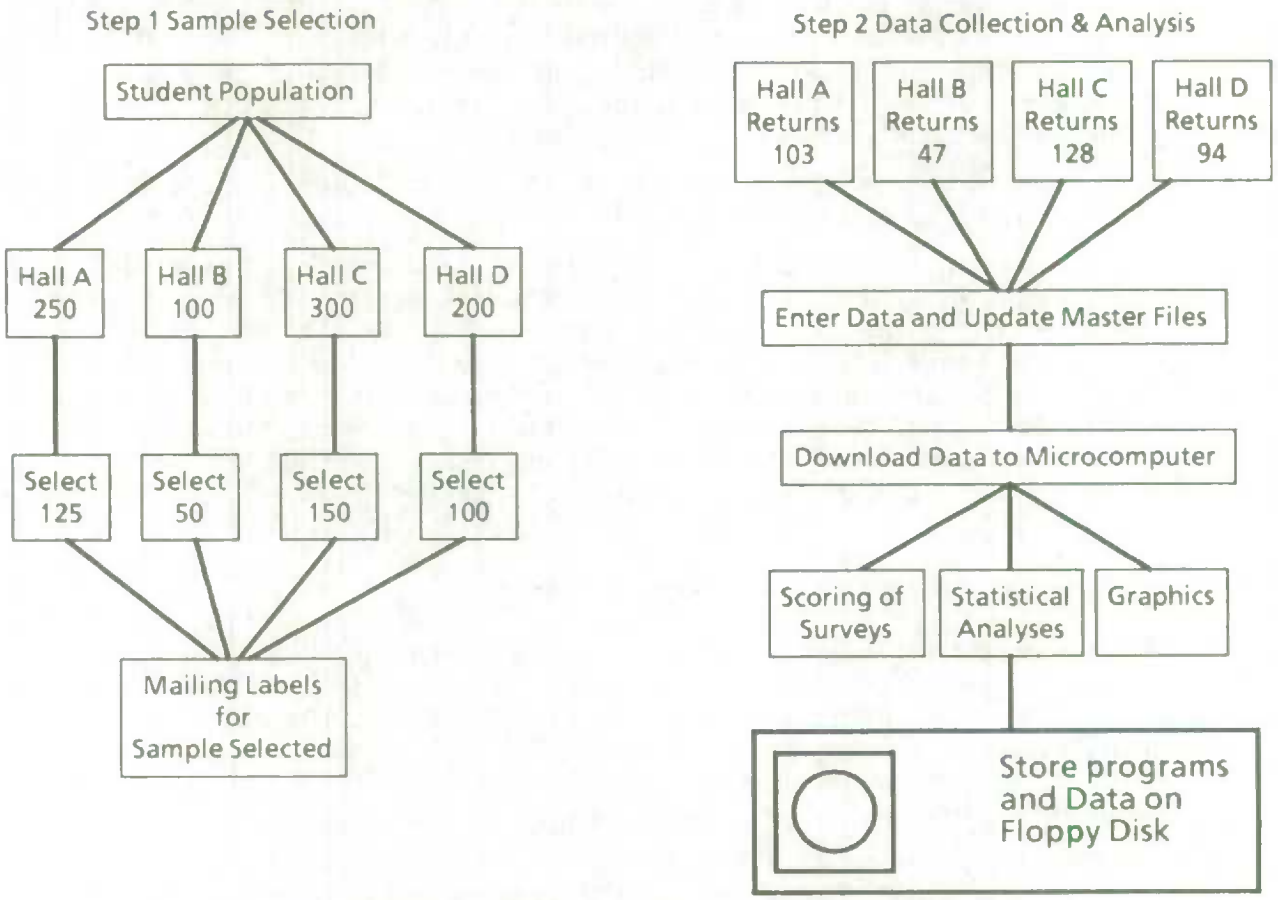


FIGURE 2 RESIDENCES ENVIRONMENT SCALE EXAMPLE

Step 1: Identification of the population to be sampled. Four residence halls (Halls A, B, C and D) were selected for examination. In the tracking system, all master files were merged to ensure that all students in those residences were in the initial population. The condition imposed to construct the correct sub-populations utilized information extracted from the current local address record. Each residence was handled separately in order to construct a stratified random sample. If required, it would have been possible to select, say, only students below a certain age, only females in a particular program of study, or only individuals who had not participated in Survey X.

Step 2: Random selection. As the tracking system constructs a sub-population of interest, it tags an observation number onto each individual, 1 to n. The tracking system's random procedure was used to generate a 50% random sample in the following way: the procedure generated a specified number of unique random numbers between 1 and n; these random numbers were then linked to the observation number tags generated in the construction of the sub-population, and those students were then identified as part of the sample. A 50% random sample was generated for each residence hall.

An update procedure was used to 'flag' these students in their master files by means of an alpha code in the field designated for the URES survey. The names and addresses for the sample were then downloaded to the PC, where SAS was used to generate labels and checklists for the mailout procedure.

Step 3: Data collection and analysis. Response time completed, data from the returned survey forms were entered on the PC. A SAS program was used to score the survey and compile the summary statistics. The residence scores, correlational analysis and plots displaying the residence scores on a number of scales, were all generated on the PC using SAS and LOTUS. The programs, worksheets, graph files and raw data were all stored on one floppy disk and filed, keeping PC memory free and collecting all the relevant programs and data in one spot for easy reference. Individual scores and respondent information were uploaded to the mainframe, where the master file survey codes were updated to indicated who had responded. The survey code then, provided in a single record field specific information regarding both who had been sampled and who responded.

8. EXAMPLE 2 FIRST YEAR MARKS DISTRIBUTION

The Learning Skills Task Group required specific course information in order to aid in the development of a new Learning Skills Centre. There was also considerable interest in the ability of new students to adjust to their new academic setting, as indicated by the average difference between their Grade 13 and first semester grades.

Information extracted from the tracking system consisted of student control number, course numbers and marks, grade 13 averages, program of study and sex. No names were required, and the control number was used to link a number of different record types.

The distributions of marks for all first year courses were generated, along with the means, standard deviations and failure rates. For this particular part of the analysis, not even the student numbers were required, just the course numbers and course marks. This analysis permitted examination of course mark distribution in a way previously unavailable except by direct application to the departments involved.

Further analysis was conducted by collecting all the course information by student number and linking this information to Grade 13 averages on the master files. The

averages change in academic grades from Grade 13 to first year university could then be compared on an individual (but anonymous) basis, and the effects of sex or program of study as possible confounding factors on this relationship were also examined.

9. FUTURE DIRECTIONS

Although the original intention of this system was to make use of available data in a tracking procedure, once the system was in place, endless possibilities presented themselves. Along with further enhancements to the menu portion of the system, future plans include updating the information selected from the Registrar's data, i.e. including pieces that now appear to be useful and eliminating those which are now outdated (eg. the OTEA scores) or are not as useful as originally anticipated.

This fall, the SESG conducted a survey of incoming students (Astin, 1987), which yielded a large amount of new information regarding the expectations and the financial and personal resources of new university students. This survey will be administered on a regular basis to the incoming freshmen, adding a new dimension to the database provided by the master files. Plans are also underway to incorporate the Post-Graduation Survey into the tracking system. Since 1976, this survey has been administered to University of Guelph graduates two years after their graduation, in order to determine how they have fared in the employment market and how they assess their university experiences. Using the tracking system, it will be possible to compare post-graduation experiences with incoming expectations and academic achievements.

In addition, there is an interest in establishing a means of regularly tracking changes of program and changes of majors within program. Simpson (1987) has documented a number of interesting uses for this information, including providing academic departments with early warning signals of program problems indicated by continual shifts in and out of departments and programs. Tracking majors is also of great interest with respect to attrition, since uncertainty regarding program of study is one indicator of dissatisfaction and potential withdrawal. Increasing awareness of the system capabilities continually prompts new requests from new client departments, and at the same time increases SESG awareness both of current concerns and the potential of the system.

10. SUMMARY

This paper has described the Student-Environment Tracking System, its data sources, its computer resources and some of its capabilities. Two examples were provided which indicated the scope and utility of the system, and future directions of the system were discussed briefly. Some of the advantages of this system and its reliance on an combination of both primary and secondary data are:

1. Accuracy of personal and academic history.
2. Discretionary use of personal and academic history in survey work while preserving anonymity.
3. Reduced respondent burden.
4. Efficient use of data by eliminating the need to collect information already gathered.
5. Automatic inclusion of variables which may not originally have been perceived as important.

6. Generation of useful and timely information without having to survey the population.

7. On-going applications for the system provided by user groups and current literature.

8. Quick sample selection and data analysis provided by flexibility of data files and computer programs.

9. Judicious use of computer facilities keeps resource costs low and guarantees a degree of self-sufficiency for a number of system capabilities.

ACKNOWLEDGEMENTS

The Student-Environment Study Group appreciates the extensive contribution of Gayle Jeffery, who wrote most of the mainframe programs for the tracking system, and who provided valuable advice during the conceptualization stages of the system's development. We also extend our thanks to Jim Burgess and Tom Rockola of the Registrar's Office, for their endless patience in explaining the details of the original data.

REFERENCES

Astin, A. (1987). The Cooperative Institutional Research Program Freshman Survey. Higher Education Research Institute, Graduate School of Education, University of California, Los Angeles CA.

Gilbert, S.N., and Gomme, I.M. (1986). Future directions in research on voluntary attrition from colleges and universities. *The Journal of the American Association of Collegiate Registrars and Admissions Officers*, 61, 227-238.

Moos, R., and Gerst, M. (1974). University residence environment scale. *Consulting Psychologists Press*, Palo Alto, CA.

Simpson, W. (1987). Tracking students through majors: methodology and applications. *The Journal of Higher Education*, 58, 323-342.

SESSION IX: CONTRIBUTED PAPERS

Chairperson: Geoff Lee, Australian Bureau of Statistics

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

AUTOMATED QUALITY ASSURANCE PROCESSING OF ADMINISTRATIVE RECORD FILES

JAMES R. JONAS and PAUL S. HANCZARYK[1]

ABSTRACT

The Census Bureau makes extensive use of administrative records information in its various economic programs. Although the volume of records processed annually is vast, even larger numbers will be received during the census years. Census Bureau mainframe computers perform quality control (QC) tabulations on the data; however, since such a large number of QC tables are needed and resources for programming are limited and costly, a comprehensive mainframe QC systems is difficult to attain. Add to this the sensitive nature of the data and the potentially very negative ramifications from erroneous data, and the need becomes quite apparent for a sophisticated quality assurance system on the microcomputer level. Such a system is being developed by the Economic Surveys Division and will be in place for the 1987 administrative records data files.

The automated quality assurance system integrates micro and mainframe computer technology. Administrative records data are received weekly and processed initially through mainframe QC programs. The mainframe output is transferred to a microcomputer and formatted specifically for importation to a prestructured LOTUS 1-2-3 worksheet. Systematic quality verification occurs within the LOTUS structure, as data review, error detection, and report generation are accomplished automatically. As a result of shifting processes from mainframe to microcomputer environments, the system eases the burden on the programming staff, increases the flexibility of the analytical staff, and reduces processing costs on the mainframe and provides the comprehensive quality assurance component for administrative records.

1. INTRODUCTION

The Bureau of the Census makes extensive use of administrative record information in our economic programs. The data originate from the business-related tax collection processes of the Internal Revenue Service (IRS) and, to a lesser extent, the Social Security Administration. During economic and agriculture censuses years, the volume of administrative record data received increases substantially. These data have enabled us to conduct economic and agriculture censuses on a timely and efficient basis and with

[1] James R. Jonas and Paul S. Hanczaryk, Economic Surveys Division, U.S. Bureau of the Census, Washington, D.C. 20233

a minimum of reporting burden on the business and farm communities. The success of our economic and agriculture programs depends to a great extent on the timeliness and quality of these administrative record files.

It is vital for Census Bureau operations to ensure the quality of all incoming data. As in past economic censuses, we have developed mainframe quality assurance programs for the administrative record data. However, since such a large number of these tables are needed and resources for programming are limited and costly, a comprehensive quality assurance system is difficult to attain entirely on the mainframe. Add to this the sensitive nature of these data and the potential ramifications of erroneous data, and the need for a more sophisticated quality assurance system becomes apparent. The Census Bureau has developed a comprehensive quality assurance system that manages various phases of our administrative records review process. This automated system will allow us to perform more thorough quality assurance within the bounds of restrictive budgets and limited programming resources.

The automated quality assurance system integrates mainframe computer and microcomputer technology. The Census Bureau has established standards that delineate our fundamental requirements of the incoming administrative record data set. These standards are entered into a microcomputer system. After the mainframe quality assurance programs are run, the results are downloaded into the same microcomputer system. The reporting patterns of the actual administrative record data are then compared to the predetermined standards. Mechanical data verification occurs as data review, error detection, and report generation are accomplished automatically at the microcomputer level. As a result of shifting processes from mainframe to microcomputer environments, the system eases the burden on the programming staff, increases the flexibility of the analytical staff, and reduces the processing costs on the mainframe. Moreover, the system provides the quality assurance component needed for thorough and unerring review of administrative records. Although designed specifically for the IRS business income tax return files used in the censuses, it can and will be adapted to all to incoming administrative record files after 1988.

2. OVERVIEW OF QUALITY ASSURANCE SYSTEM FROM A MANAGEMENT PERSPECTIVE

Administrative records play a major role at the Census Bureau, a role that has steadily grown in importance over time. The increasing need for more and better statistics, the need to compile those statistics with a minimum of burden on the private sector, and the need to use our available human and financial resources as efficiently as possible have all contributed to the importance of administrative records.

Over the past several years, the quality of the administrative records generally has been excellent. However, we did experience certain problems with the quality of the 1982 business income tax data from the IRS. The most detrimental problem was the inadequate quality of the principal industrial activity codes for sole proprietorships. As a result of this problem, the Census Bureau published only limited statistics for nonemployers in the 1982 Economic Censuses. If our quality assurance programs had been more sophisticated, the errors could have been identified earlier and the negative impact would have been minimized.

Heading into the 1987 Economic Censuses, it was determined that additional measures were needed to ensure the quality of administrative record data received from the IRS. An overall quality management system responsive to certain factors that have adversely affected past administrative data sets was necessary. The three major factors that have plagued us in the past are:

1. **Vast amounts of administrative record data**

The IRS will provide us with selected business 1987 tax return data (received in 1988) for various legal forms of businesses, including corporations, S corporations, foreign corporations, partnerships, nonprofit organizations, and sole proprietorships. In total, the Census Bureau expects over 75 million tax return records in 1988. Attachment 1 details the approximate number of administrative records that will be used in the 1987 Economic and Agriculture Censuses for the various form types. Clearly, the number of data records received during census years is immense, but the complexity of the required quality assurance goes beyond sheer volume. A data record often contains several data items, each greatly increasing the detail of the individual records and the entire data files. Moreover, not all form types contain the same set of data items, nor do they have the same pattern of receipt. Consequently, in addition to performing quality review for over 75 million individual records, the Census Bureau must also be concerned with assuring the quality of the various data items on those 75 million records. Attachment 2 details the contents of the 1987 business income tax return records for each of the different form types.

Additionally, businesses file their tax returns with one of ten IRS centers. Each of the individual centers processes the returns, and the quality of data received from different service centers can vary. The Census Bureau reviews data at the service center level in response to such variation.

2. **Restrictive budgets**

Restrictive budgets are another major factor that contribute to the difficulty of assuring the quality of the administrative record data. In keeping with the overall governmental policy on spending, the Census Bureau is attempting to provide greater services at less cost. Workloads for programming staffs increase significantly during census years, yet the staffs do not expand proportionately. The quality assurance processing, which relies considerably on various computer resources, can be adversely affected. It is also important to note that most quality assurance processing is traditionally done at the mainframe computer levels. Use of the Census Bureau's mainframe computer is costly and becomes more so as increasingly larger data files are processed.

3. **Lack of communication between agencies**

Miscommunication or lack of communication between agencies has contributed to past administrative record problems. Clear lines of communication between the Census Bureau and the agency providing the data during all phases of the procurement process also are essential for assured data quality. The agencies first must agree upon the data files and the specific data items that are needed and that can be provided. Certain data that the Census Bureau requests may not be available or in some cases affordable. Any discrepancies must be resolved in time to avoid delays, which could affect data utility. Moreover, the agencies must agree upon the expected quantity and quality of the administrative data. Requirements that quantify the Census Bureau's expectations of the incoming data should be established.

The development and implementation of the quality assurance system represent a comprehensive response to the administrative record data problems we encountered in the past. The system provides for the review of large and complex IRS data files, promotes frequent interagency communication, and identifies errors instantly. The major element of the quality assurance system is the mechanized data verification. Basically, the Census Bureau establishes standards that detail our fundamental requirements of the

incoming IRS data. The reporting patterns of the actual data are compared to these standards, and systematic data verification occurs at the microcomputer level. The Census Bureau then prepares status reports indicating whether the data conforms to the standards.

Census Bureau staff members develop the standards far in advance of the actual receipt of the data. This gives the IRS ample opportunity to examine the requirements for reasonableness and request adjustments if necessary. The requirements are divided into timing standards and quality standards. The timing standards list the estimated total number of tax returns for the different types of businesses and the estimated number to be received by various dates. The quality standards detail the expected reporting patterns of specific data items.

The mechanized data verification technique simplifies our analytical review process. A series of results tables are created that compare the actual data to the expected standards. Discrepancy flags are set for those data components that do not meet the standards. This approach minimizes the risk of analytical omissions during the review process.

Status reports comparing the reporting patterns of actual data to the predetermined standards are sent to the IRS monthly. These status reports are a subset of the comprehensive results tables, detailing only the basic requirements of the IRS data set. The status reports promote communication between the agencies. If data problems exist, they are illustrated in the report. Immediately, the Census Bureau and the IRS must decide upon any remedial action or recovery efforts necessary to prevent compromising the censuses. Timeliness is crucial because the IRS data tapes are not kept indefinitely. If errors are not identified early and remedial action is not implemented in time, recovery of the data may not be possible or may become extremely costly.

The quality assurance system is not designed to guarantee that administrative data problems will never occur. It does serve, however, to document our requirements formally so that the characteristics of the data set are not left to chance, and monitoring and early error identification are possible.

3. DETAILS OF AUTOMATION

Administrative record data files are received weekly and processed initially through mainframe quality assurance programs. The mainframe programs are prepared well before the administrative data files are received and generate the initial quality assurance tables that are fundamental to the entire review process. Traditionally, mainframe programmers were responsible for creating the entire data tables, which included data cells and the surrounding text (i.e., headers and stubs). However, for the data table programs associated with the 1987 Economic Censuses, the two data table components are handled separately. Data tabulation is performed as usual at the mainframe level whereas table text is created at the microcomputer level by nonprogrammers. A procedure has been developed that generalizes data tables for all administrative records files. This procedure has allowed the Census Bureau to design a microcomputer program that is capable of building table images for any administrative records file. Once built, the table images are uploaded to the mainframe and used by programmers to align data tabulation files. The job of programming the quality assurance tables is greatly simplified, as table image formation is handled by nonprogrammers, leaving mainframe programmers adequate time to concentrate their efforts solely on data tabulations. Attachment 3 illustrates one of the various mainframe table that is produced for each of the different forms of organization. This table shows the weighted distribution of Form 1040, Schedule C records by service center by net receipts size class.

The mainframe computer performs only the basic data tabulations of the administrative records files (i.e., generates current tables). The output from these mainframe quality assurance programs is downloaded to a microcomputer, and all remaining review operations are automated at the microcomputer level. The various operations performed on the microcomputer include calculating percentages used in the review of the current tables, producing cumulative tables, performing key data item verification, and generating quality assurance status reports. Developing this systematic approach, using mostly microcomputer technology, has allowed greater flexibility of review as well as lessened the workload of mainframe programmers.

The mainframe quality assurance output is imported into a prestructured LOTUS 1-2-3 worksheet on the microcomputer. This worksheet also will contain the predetermined standards that outline the Census Bureau's expectations of the incoming data set. Automatically, a mechanical table review and data verification are performed; and inconsistencies between the actual data sets and the standards are identified within the results tables. The two major benefits of this data verification system are:

1. It enables us to easily spot problems in the data. Data components that do not meet the standards are flagged for analyst review. The possibility of overlooking errors in the administrative data is minimized.

2. It directs us to areas of the data that require further investigation. The results tables oftentimes lead us to problems even though the overall standards are met. For example, certain unexpected trends in the results report are reviewed in additional detail. In effect, the results tables enable us to concentrate on those areas that may contain problems. This may involve additional review at the service center level, or it may even require us to download records with these certain characteristics to the microcomputer. We then review these records on a manual basis in an effort to spot the problem.

As previously stated, the standards detail the basic data quality requirements that are essential to the 1987 Economic and Agriculture Censuses. This procedure of automatic quality verification (i.e., comparing the incoming data to predetermined standards) allows us to determine immediately if the basic quality of the incoming data is acceptable.

After current cycle review and verification, cumulative tables are prepared on the microcomputer. This technique of producing cumulative tables on the microcomputer rather than the mainframe provides a more efficient use of our resources. First, it eliminates the need to retain cumulative files on the mainframe system, which reduces computer costs. In the past, these cumulative files were retained on the mainframe and added to each subsequent current cycle to form the next set of cumulative tables. Using microcomputers, simple formulas were established within LOTUS 1-2-3 that created cumulative tables at virtually no cost. Secondly, the quality assurance tables for the cumulative portion do not require mainframe programming. A printout of the cumulative quality assurance tables are produced and retained for analysis and documentation purposes.

In addition to this comprehensive set of cumulative tables, we produce a set of results tables. As was the case with the current cycle, these results tables detail comparisons of certain key data items. Attachment 4 shows one of the many results tables that is produced for the cumulative quality assurance. This table details the actual number and percent of the weighted Form 1040, Schedule F records by service center, together with the expected percent. As can be seen, the cumulative data are reasonable and fall within the acceptable standards. If inconsistencies did exist, the applicable service center would have been flagged. The final component of the automated quality review process is the generation of a report detailing the status of the cumulative IRS data file. This report

compares the overall quality of the data set to the expected quality indicated in the timing and quality standards. The reports are generated and provided to the IRS approximately monthly. As discussed earlier, the status reports capsulize the quality of the administrative data for representatives of both agencies, which promote frequent interagency communication.

4. RESULTS OF QUALITY ASSURANCE REVIEW

The timing and quality status reports can serve to alert both the Census Bureau and the IRS of data problems in their early stages and facilitate cooperative action by both agencies. In most of the cases, however, the timing and quality standards alert us of changes in respondent reporting patterns. These circumstances require no corrective action by the IRS, but they may have cost and processing implications for the Census Bureau in the 1987 Economic and Agriculture Censuses. Attachment 5 illustrates this point well. Through late May 1987, the Census Bureau had received approximately 697,600 Form 1120 returns (i.e., corporations) with a standard of 760,000 returns. The standard for the number of Form 1120 returns was not met. However, the shortfall in the number of Form 1120 returns was offset by an increase in the number of Form 1120S returns (i.e., S corporations). The Census Bureau had received approximately 328,850 Form 1120S returns, far exceeding the standard of 225,000. The shift in the number of returns for these two types of corporations resulted from the perceived advantages in the new tax law associated with filing Form 1120S rather than Form 1120. Although this represented a legitimate shift in taxpayer reporting patterns that was not a data error, the information was pertinent to our processing. We are implementing a procedure for 1987 that will account for such a shift from corporations to S corporations. Attachment 6 illustrates one of the various tables from the quality portion of the report. As indicated, the quality of these data meets the standards for each of the basic data items. If an item had failed the standard, it would have been flagged for analyst research.

The automated quality assurance of administrative records files will be completely operational for the 1987 IRS data files. Prototypes of the system have been and are being used for the 1985 and 1986 IRS business income tax files. For both years the automated process and the entire quality assurance system have been instrumental in the successful procurement and review of the IRS data files received for the censuses.

The integration of both mainframe and microcomputer technology in the automated quality assurance system has allowed the Census Bureau to effectively and comprehensively assure the quality of the large data files provided by the IRS. In addition, mainframe computer programmer workloads have been and will continue to be lessened since much of the automation was designed and is controlled by nonprogramming staff and is implemented in a microcomputer environment. Mainframe computer resources are reduced and programming burden is lessened allowing programmers to concentrate their efforts on basic data tabulation. Also important, the automated system provides the flexibility of review for different levels of personnel. Managers can review the summarized timing and quality report and determine the status of the business income tax files quickly and efficiently. Subject-matter analysts will review the more comprehensive quality assurance reports that are produced weekly. As mentioned above, the quality assurance system will direct the analysts to the data elements that require further investigation.

5. SUMMARY

The Census Bureau has designed an overall quality assurance system that is comprehensive and responsive to the potential problems and limiting factors of complete quality assurance. The system responds to the large volumes of IRS data by interacting with the IRS closely and promptly to ensure proper data procurement. The expected quality of these large data files is jointly determined and agreed upon with the IRS through the timing and quality standards and is verified by the automated QC process. Given this automated process, data verification can occur within the bounds of restrictive budgets and limited programming resources. Microcomputer technology has increased the role and flexibility of subject-matter analysts while lessening the burden of mainframe programmers. Communication with the IRS is frequent and productive, resulting in efficient procurement procedures and improved data quality awareness on the part of IRS and the Census Bureau as well. This collective response to past difficulties will ensure the Census Bureau of receiving the data necessary to conduct the 1987 Economic and Agriculture Censuses in the best manner possible.

Attachment 1

The Approximate Number of Administrative Records Used in the 1987 Economic and Agriculture Censuses for the Various Form Types by Tax Year

Type of Record	Numbers of Records		
	1985	1986	1987
Business Income Tax Files	2,617,000	20,051,000	30,881,000
Form 1040, Schedule C	--	11,750,000	12,500,000
Form 1040, Schedule F	2,450,000	2,450,000	--
Form 1040, Schedule SE	--	--	10,000,000
Form 1120	42,000	2,550,000	2,650,000
Form 1120-A	--	200,000	210,000
Form 1120F	--	11,000	11,000
Form 1120S	17,000	900,000	950,000
Form 1065	108,000	1,750,000	1,800,000
Form 990	--	380,000	400,000
Form 990-PF	--	35,000	35,000
Form 990-T	--	25,000	25,000
Form 1120S, Schedule K-1	--	--	700,000
Form 1065, Schedule K-1	--	--	1,600,000
Annual Tax Files:	41,950,000	43,500,000	45,050,000
IRS Business Master File	24,000,000	25,000,000	26,000,000
IRS Payroll and Employment File	17,000,000	17,500,000	18,000,000
SSA Business Birth File	950,000	1,000,000	1,050,000
Total	44,567,000	63,551,000	75,931,000

Attachment 2
Contents of 1987 Business Income Tax Return Records for the Various Form Types

Data Item	Form Type								
	1040 Schedule C	1040 Schedule SE	1065	1120 & 1120-A	1120 S	1120 F	990	990-PF	990-T
Taxpayer name and last name character counts	X								
Taxpayer mailing address and city character counts	X								
Taxpayer SSN	X	X							
Spouse SSN	X								
EIN	X		X	X	X	X	X	X	X
PIA	X		X	X	X	X			
Gross receipts or sales	X		X 1/	X 1/	X 1/	X 1/			X 1/
Returns and allowances	X								
Accounting period	X	X	X	X	X	X	X	X	X
Wages	X								
Cost of Labor	X								
End-of-year code	X		X		X				
Months actively operated	X		X		X				
941 indicator	X								
Gross royalties				X					
Reported total gross receipts							X		
Total receipts								X	
Last names and SSNs (up to ten partners) 2/			X						
Last, names, SSNs , and number of shares of stock (up to ten shareholders) 2/					X				
Net earnings from farming		X							
Net earnings from nonfarming		X							
Total net earnings from self-employment		X							

1/ Less returns and allowances
2/ From Schedule K-1.

Attachment 3
Weighted Distribution of 1040 Schedule C Records by Service Center by Net Receipts 1/ Size Class

Service Center	Total	Net Receipts Size Class										
		<0	Blank or 0	1-2,499	2,500-4,999	5,000-9,999	10,000-24,999	25,000-49,999	50,000-99,999	100,000-249,999	250,000-499,999	500,000 +
All Centers												
Atlanta												
Philadelphia												
Austin												
Cincinnati												
Kansas City												
Andover												
Ogden												
Brookhaven												
Memphis												
Fresno												
Others												

1/ Gross receipts less returns and allowances.

Attachment 4
Percent of Weighted 1986 Form 1040, Schedule F Records by Service Center

Tax Year	Total Schedules	Services Centers										
		Atlanta	Phila-delphia	Austin	Cin-cinnati	Kansas City	Andover	Ogden	Brook-haven	Memphis	Fresno	Others
1986												
Count	2,087,200	176,700	71,600	374,900	262,100	358,600	118,800	343,200	40,300	288,100	52,500	400
Percent	100.0	8.5	3.4	18.0	12.6	17.2	5.7	16.4	1.9	13.8	2.5	0.0
Expected Percent	100.0	8.5	3.0	18.5	11.5	17.5	5.5	16.5	2.0	14.0	2.5	0.0
Expectation 1/ Not Satisfied												

1/ Acceptance interval of + or - 2.0 percent.

Attachment 5

The Weighted Number of 1986 Form 1120 Returns by Various Dates

Date	Form 1120 Returns		Requirement Not Satisfied
	Actual	Required	
Late March 1987	326,500	303,000	
Late April 1987	697,600	760,000	Not Satisfied
Late May 1987		988,000	
Late June 1987		1,190,000	
Late July 1987		1,418,000	
Late August 1987		1,621,000	
Late January 1988		2,077,000	
Late October 1988		2,533,000	

The Weighted Number of 1986 Form 1120S Returns by Various Dates

Date	Form 1120S Returns		Requirement Not Satisfied
	Actual	Required	
Late March 1987	103,350	90,000	
Late April 1987	328,850	225,000	
Late May 1987		292,000	
Late June 1987		352,000	
Late July 1987		420,000	
Late August 1987		480,000	
Late January 1988		615,000	
Late October 1988		750,000	

Attachment 6

Data Element Reporting Patterns of Weighted 1986 Form 1120S Returns

Data Elements	Percent of Form 1120S Returns		Requirement Not Satisfied
	Actual	Required	
EIN			
Blanks, all zeros, or nonnumerics	0.0	Less than 1.0	
Invalid IRD	0.0	Less than 1.0	
P8A CODE			
Blanks or nonnumerics	0.0	Less than 6.0	
Blanks, nonnumerics, unclassified, or invalid PBA codes	11.5	Less than 18.0	
GROSS RECEIPTS OF SALES LESS RETURNS AND ALLOWANCES			
Blanks, all zeros, or nonnumerics	20.9	Less than 40.0	
Of records with a positive numeric entry, the percent in various size ranges:			
- Less than $100,000	45.7	30.0 - 60.0	
- Greater than or equal to $100,000 and less than $500,000	36.9	20.0 - 50.0	
- Greater than or equal to $500,000	17.4	10.0 - 30.0	
ACCOUNTING PERIOD			
Blanks, all zeros, or nonnumerics	0.0	Less than 1.0	

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

AN ALGORITHM FOR THE DETERMINATION OF OPTIMUM MATCHING RULES FOR THE LINKAGE OF RECORDS FROM TWO SOURCES

YASAR YESILCAY[1]

ABSTRACT

An algorithm is presented which searches in a stepwise fashion optimum rules for linking records from two sources. The rules found are optimum in the sense that the resulting net matching error or matching bias (B), and the gross matching error (G), satisfy the following inequalities:

$$0 \leq G \leq g \text{ and } |B| \leq b$$

where b and g are some pre-specified maximum levels for the two errors B and G respectively, and $|B|$ is the absolute value of the bias.

1. INTRODUCTION

A definition of record linkage is given by Dunn (1946) in the following poetic words:

Each person in the world creates a Book of Life. This Book starts with birth and ends with death. Its pages are made up of the records of principal events in life. Record linkage is the name given to the process of assembling pages of this Book into a volume.

When "person" in the above definition is interpreted as a population item, whether it is a population of events or people or things, then the definition covers the current usage of the term in a large number of fields. The "Book" may have as few as two "pages" (or records) and this paper is concerned with such cases, i.e., the linkage of records from two different sources. The procedure developed here, may of course be extended (with appropriate changes in the definitions of some concepts involved) to the case of record linkage from many sources. Linkage of records from two or more sources is applied in a large number of fields such as business, history, demography, sociology, education, travel and information retrieval, to name just a few.

If a unique and permanent identification code can be given to each and every element in the population (or sample) under study and such a code is reported without any error on every record generated for these elements, then there would not be a need to search for any matching rule. The records would then be compared on this code and those records that have the same code will be declared as belonging to the same element and hence will be linked (or matched); the remaining records will then be declared as unmatched. Such **ideal** (perfect) identifying information must have the following features: unique, permanent, available and highly discriminatory (Nitzberg and Sardy 1965). In most applications, however, the identifying information that can be used lacks some or all of these features.

[1] Yasar Yesilcay, Sultan Qaboos University, College of Science, P.O. Box: 32486 AL-Khoudh, Muscat, Sultanate of Oman.

Using imperfect information for record linkage may result in three types of linkage (or matching) errors: erroneous matches, erroneous nonmatches and mismatches. When a record from a given source does not have a counterpart in the other source but is declared as matched to a record in the latter, such a record is said to be an erroneous match. Similarly, if a record from a given source has a counterpart in the other but is declared as unmatched, then there is an erroneous nonmatch. Finally, if a record has a matching counterpart in the other source but is matched with a different record, then there is a mismatch. (See Marks, Seltzer and Krotki, 1974 or Yesilcay, 1975 for a detailed description and discussion of these errors.)

Rules for record linkage operations are determined to minimize such errors when only imperfect information is available. However, it is an established fact that too "tight" rules while giving fewer erroneous matches and mismatches will result in too many erroneous nonmatches. On the other hand, too "loose" rules give fewer erroneous nonmatches but too many erroneous matches and mismatches. Thus a balance is needed between the two extremes. Such a balance is achieved by using a set of "optimum" rules, where the definition of optimum depends on how the results of linkage will be used and how the linkage errors influence that use.

Sunter and Fellegi (1967) proposed an approach for obtaining optimum rules where all possible pairs of records, one from each source, are compared and each pair is classified as matched or unmatched or possible match. Then, optimum matching rules is defined as that set of rules which minimizes the probability of positive disposition of records for some pre-specified error rates. The theory behind this approach was later revised by Sunter (1968) and further refined by Fellegi and Sunter (1969).

Tepping (1955, 1960, 1968) developed a record linkage approach for subcription fulfillment which he later (1969) adopted for a dual record system. His optimum rules are aimed to minimize the cost of erroneous matches and erroneous mismatches, and require the unit costs to be given. It was argued (Yesilcay, 1975) that imputation of these costs is not practical in a dual record system although it was quite easy for the subscription problem. Also, Tepping treats mismatches as erroneous matches, whereas in a dual record system mismatches have no influence on the estimate of the total number of events or on the variance of the estimate.

For dual record systems, Marks, Seltzer and Krotki (1974) have defined optimum matching rules as that set which yields an equal number of erroneous matches and erroneous nonmatches, thus giving zero net matching error or bias (= B= number of erroneous matches - number of erroneous nonmatches) with a minimum gross matching error (= G = number of erroneous matches + number of erroneous nonmatches). Although these criteria yield an unbiased estimate of the total number of events, Yesilcay (1975) has shown that the variance of the estimate is influenced by the gross matching error and a biased estimate may prove to be better in terms of the mean square error of the estimate in some cases. Thus, the criteria of optimality are changed and optimum matching rules are defined as that set of rules which satisfy the following two inequalities:

$$0 \leq G \leq g \text{ and } |B| \leq b$$

where b and g are pre-specified maximum error levels; B and G are the net and gross matching errors as defined before. One has to keep in mind that the values of B and G differ depending on in which source one counts the erroneous matches and erroneous nonmatches. It is suggested that these are counted in the source that has the minimum number of records.

A set of axioms are needed for the algorithm explained in this paper which yields the optimum rules as defined above. These axioms are given in the next section.

2. AXIOMS NEEDED

The search routine designed to determine the optimum record linkage rule is based on the following axioms:

1. Records from two sources are available; for each element of the population or sample, at most one record exists from each source.
2. These records have information on some variables that can be used for decision on whether or not a pair of records, one from each source, refer to the same element. Such information will be called matching information.
3. The matching information is imperfect, in the sense that it lacks one or more of the desirable properties given in the previous section.
4. When comparing the records from the two sources, some matching (or linkage) errors are usually unavoidable, and although not desirable, such errors can be tolerated.
5. Increasing the number of variables on which the records are simultaneously compared decreases the number of erroneous matches but increases the number of erroneous nonmatches in both sources.
6. Although "exact agreement" on each variable is desirable in order to minimize the erroneous matches and to simplify the matching operations, it is sometimes preferable to use agreement within some specified (looser) tolerance, since this will reduce the number of erroneous nonmatches. However, a too "loose" level of tolerance will also result in too many erroneous matches and mismatches and hence is not desirable, i.e., for each variable there exists an **"optimal tolerance level of agreement"**, which in turn depends on other variables that are used for linkage and their respective optimum tolerance level of agreement.
7. For determining the matching rules, a sample of records from each system is available for which the **true match status** of each record is known. These records may come from a pretest or from the initial stages of an actual operation. It is important that the quality of information on these records be representative of the quality of the records that will be matched using the rules determined. For this reason a random sample of records from those of an actual operation should be preferred whenever possible and the rules found should be periodically tested when there is reason to believe that the conditions of the operation have changed.

It is obvious that many of these axioms are needed for any approach or algorithm that searches matching rules (optimum or not). A procedure for determining the true match status of a sample of records is explained in the next section.

3. TRUE MATCH STATUS OF A SAMPLE OF RECORDS

For deciding on the true match status of all the records in the sample that will be used for determining the matching rules, the procedure suggested by Marks, Seltzer and Krotki (1974) and summarized below is recommended.

In this procedure, three or more experts, or groups of experts, working **independently** classify each and every record as matched or unmatched, by their own implicit matching rules. Madigan and Wells (1974), reporting their experience in using such a procedure for the Philippines, have stated that,

"At this stage ..., all information on the records, any knowledge of cultural characteristics, all expertise upon interviewing procedures and all other germane information and knowledge should be utilized ..."

as part of the implicit rules. Once a decision is reached on all records by the three experts, the next step is to compare their decisions for each record. If they all agree that a pair of records is a match then that is accepted as the "true" match status of the pair. Similarly if they all agree that a record does not have a counterpart in the other source, then it is accepted as a "true" nonmatch. In case of disagreement, all experts are asked to reconsider their decisions. If they still disagree, then the record is sent back to the field for reverification and additional information. Also, the nonmatches, and perhaps a sample of the matches are sent to the field to check for conceptual, temporal and geographic out-of-scope events and for improved information. After these, the records are reconsidered and a decision is reached based upon the additional information, thus classifying every in-scope record as a "true" match or a "true" nonmatch.

It is obvious that the above procedure, which results in the best decision under the given conditions for the match status of each and every record (with hopefully no errors) can not be used every time such a matching operation is to be performed, especially when this operation is to be repeated a large number of times or when the number of records to be matched is very large. The true match status of the records in the small sample are determined by the above procedure so that these records can be used for determining the matching status of all records that are and/or will be involved.

Now one is ready to determine a set of explicit rules that will reproduce the match status of the records in the sample in a less expensive, laborious and time consuming way while satisfying the criteria of optimality for such rules.

Marks, Seltzer and Krotki (1974) give details for such a procedure where the criteria for optimality are zero net error and minimum gross error. Although such rules are desirable under certain conditions, it was observed that their achievement is not always possible (Yesilcay, 1975). Furthermore, the estimated number of events found by this procedure, though unbiased, may have a larger variance due to G, than a biased estimate.

The algorithm given in this paper achieves optimality in a slightly different sense with $0 < G \leq g$ and $|B| \leq b$ where b and g are predetermined error levels to obtain the minimum mean square error of the estimated total number of events in a dual record system. For other applications, these values may be predetermined to achieve other goals. Using this algorithm, one may still achieve, if feasible, the optimum defined by Marks, Seltzer and Krotki (1974) simply by setting b equal to zero.

The procedure which determines the matching rules to satisfy the above stated criteria is called the Steps Approach to Matching (SAM) and is explained in the next section.

4. THE STEPS APPROACH FOR DETERMINING OPTIMUM MATCHING RULES

This approach determines optimum matching rules in two or more steps (or stages). In this way one does not have to look at all possible tolerance levels of all possible combinations of the variables to achieve the desired (predetermined) bounds on errors, since the number of such combinations is extremely large even for a computer to handle within reasonable time limits.

In the first step, as many of the true matches as possible are identified and removed from the pool of records with a minimum number of erroneous matches. This is done by matching the records of the two sources on all possible combinations of a small number of variables (3 to 5) at their "exact" tolerances or some fixed "tight" tolerance level. Among

the combinations that yield the same minimum erroneous matches, the algorithm selects that combination with the maximum number of true matches. Erroneous nonmatches of this step may be large and the purpose of Step II is to reduce this to a reasonably low level consistent with the predetermined error levels.

In Step II only the records that have been classified as unmatched in the previous step are compared. To reduce the cost of searching and application the search is limited to all possible 2 variable combinations in Step II. Furthermore, the maximum number of "reasonable" levels of tolerance for each variable is limited to five. Finally, the total number of variables that can be used for the matching operation is limited to 20. When there are more than 20 such variables, one has to choose the most discriminating twenty of them. Marks, Seltzer and Krotki suggest using the gross matching error as a proxy for the discriminating power of a variable. Yesilcay (1975) proposes the modal frequency of the variable as a poxy for the discriminating power, where a low modal frequency indicates a high discriminating power.

A novel feature of the SAM Algorithm is its control on the upper limit of the gross error. The search process eliminates in Step I those combinations that are likely not to meet that criterion. In Step II each combination is again checked for that criterion and such a combination is not considered as an alternative if the sum of erroneous matches of the two steps and the erroneous nonmatches of Step II is greater than g, the specified limit on gross error. Also, if in Step I this criterion is met by using 3 variable combinations, then 4 and 5 variable combinations are not searched in an effort to reduce the cost and to simplify the process of matching. Similarly, the 5 variable combinations are searched only if the criterion can not be satisified with 3 or 4 variable combinations. The operational stages of Step I and Step II are given in Table - 1.

5. CONCLUDING REMARKS

A slightly different version of the program was tested and found satisfactory using data collected in a dual record system in the Philippines (Madigan and Wells, 1974 and Yesilcay, 1975.)

A listing of the program can be obtained from the author.

Table 1

The Operational Stages of the Steps Approach to Matching (SAM)

STEP I

1. Establish the "true" match status of a sample of records.
2. Specify the number of variables that can be used for matching (NVAR $\leq$ 20), the code for the source where the errors will be counted (ISMALL), the maximum gross error, g and the absolute value of the maximum net error, b.
3. Set the number of variable combinations, NV = 3.
4. Match at "exact" tolerance levels of all NV variable combinations.
5. Choose the combinations that yielded the minimum number of erroneous matches and the maximum number of true matches.
6. If the minimum number of erroneous matches is $\leq$ g/4 then go to Step II. The combination of variables that gives this result will be used in Step I of the matching operation.

7. If the minimum in number of erroneous matches > g/4, set NV = NV + 1 and go to 4.

REPEAT 4 TO 7 UNTIL MINIMUM NUMBER OF ERRONEOUS MATCHES IS $\leq$ g/4, or NV > 5.

STEP II

1. Use the records of the previous step that are declared unmatched.
2. Match at all possible combinations of all reasonable levels of all **two variable combinations.**
3. Among all such combinations choose the one that yields a gross error $\leq$ g and an absolute bias $\leq$ b. The combination of variables together with their tolerances that gives this result will be used in Step II of the matching operation.
4. If none of the combinations tested above satisfy the requirements, then repeat Step II on the nonmatches of this step. That will be Step III. Proceed to Steps IV, V, etc., until the desired results are reached.

REFERENCES

Dunn, H.L. (1946). "Record Linkage." *American Journal of Public Health* 36, 1412-1416.

Fellegi, Ivan P., and Sunter, Alan B., (1969). "A Theory for Record Linkage." *Journal of the American Statistical Association*, 64, 1183-1210.

Madigan, Francis C., and Wells, H. Bradley (1974). "Report on Matching Procedures of a Dual Records System in the Southern Philippines." Unpublished Report.

Marks, Eli S., Seltzer, W., and Krotki, Karol J. (1974). *Population Growth Estimation, A Handbook of Vital Statistics Measurement.* The Population Council, New York.

Nitzberg, David M., and Sardy, Hyman (1965). "The Methodology of Computer Linkage of Health and Vital Records." *Proceedings, Social Statistics Section, American Statistical Association*, 1963, 100-106.

Sunter, Alan B. (1968). "A Statistical Approach to Record Linkage." in E.D. Acheson, editor, Record Linkage in Medicine, *Proceedings of the International Symposium Oxford July 1967*. London, E & S Livingstone, Ltd.

Sunter, Alan B., and Fellegi, Ivan P. (1967). "An Optimal Theory of Record Linkage." *Bulletin of the International Statistical Institute, Proceedings of the 36th Session, Sydney, 1967*, Vol. XLII, Book 2, 809-838.

Tepping, Benjamin J. (1955). *Study of Matching Techniques for Subscription fulfillment*. Philadelphia, National Analyst Inc.

Tepping, Benjamin J. (1960). *Progress Report on the 1959 Matching Study*. Philadelphia, National Analyst Inc.

Tepping, Benjamin J. (1968). "A Model for Optimal Linkage of Records." *Journal of the American Statistical Association 36*, No. 324, 1321-1332.

Tepping, Benjamin J. (1969). "The Application of Linkage Model to the Chandrasekar-Deming Technique for Estimating Vital Events." Paper prepared for the Population Council Seminar on Optimum Procedures for Matching Two Lists of Vital Events, New York, April 17, 1969.

Yesilcay, Yasar (1975). The Mean Square Error of the Estimate as a Criterion for the Assessment of Alternative Approaches to Matching in a Dual Record System. Institute of Statistics Mimeo Series, No. 1035, The University of North Carolina at Chapel Hill.

SESSION X: INVITED PANEL DISCUSSION

Chairperson: G. Brackstone, Statistics Canada

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

NOTES FOR A PANEL DISCUSSION ON STATISTICAL USES OF ADMINISTRATIVE DATA

JOHN W. GRACE[1]

My role as a member of this panel is both humbling and daunting: that of representing your subjects in Canada, 25 million with our often disorganized, sometimes messy, and frequently just humdrum lives; representing both our apprehensions and indifference; representing, for example, a young man with a nursing degree, who graduated from university, who tested positive for tuberculosis, is divorced and the proud owner of 1.9 children, a microwave and indoor plumbing.

(For those of you who don't recognize my subject, he is a composite of several Statistics Canada's databases listed in the Personal Information Index.) And he, I am certain, would not know what administrative data is, let alone what you are doing with it. As a matter of fact, neither am I. Certainly, when I was asked to speak to you, I was not even sure what was encompassed by the term "administrative data".

But I did know that the **Privacy Act** uses the term "administrative purpose" and it is defined as follows:

> "administrative purpose, in relation to the use of personal information about an individual, means the use of that information in a decision-making process that directly affects that individual".

Individuals have clearly spelled out rights of access to such information. Administrative data represent a twilight zone.

I don't want to appear presumptuous in speaking for the nurse, or any of those other 25 million Canadians. Most can and do speak very well for themselves when, for example, they are offended by survey questions, and growing numbers are. Yet many still do not care one way or another what Statistics Canada knows about them. They are the ones who have "nothing to hide" -- a rather boring confession, it has always seemed to me.

But I do speak for Canadians because Parliament has put in place a **Privacy Act** which sets out the ground rules for the federal government's collection and use of personal information. It has given the Privacy Commissioner responsibility for investigating complaints and overseeing government compliance with the act. And I have learned enough about the subject of your meeting to conclude that surveys of administrative data, being as they are indirect, effectively deprive Canadians of an opportunity to have their say for they simply don't know how their personal information may be being used.

Let me put my comments in this perspective. As some of you may have heard me say at an earlier conference, I recognize Statistics Canada's sensitivity to privacy. It has demonstrated its commitment by creating an Access and Privacy Board to review sensitive or contentious requests. And I know that commitment to the principle of individual privacy is shared by statistical agencies among the Western democracies.

[1] John W. Grace, Privacy Commissioner of Canada, 112 Kent, Ottawa, Ontario. K1A 1H3

Guy Labossière and I had the privilege of attending a conference organized by **Statistics Sweden** in June, a meeting which brought together statisticians and data protectors in common cause, not as enemies.

But this shared sensitivity has not prevented governments from linking and matching its administrative data bases to draw up, for whatever good reasons, sufficiently detailed profiles on individuals to send chills down the spine of the most tolerant of data protectors.

The statistical uses of administrative data is part of our steady drift toward data surveillance. As a society we find video, audio or medical surveillance abhorrent. But Orwell's Big Brother pales beside the potential invasiveness of linking and mining various data bases to draw up profiles on unsuspecting individuals or groups of individuals.

Of course, many Canadians have already resigned themselves to losing control of their information once it is out of their hands. They may not like what is done with it -- if indeed they even know what is done with it--but they appear disquietingly resigned. I think that unhealthy for society.

Statisticians and researchers are beneficiaries of this resignation. After all, a survey of administrative records has a 100 per cent response rate, unambiguous answers, lower administrative costs and less aggravation than other surveys. But where is it all going to end? How much of my life is it reasonable for governments to know in the interests of research or social planning? I do not believe that this fundamental question has been sufficiently addressed by either statisticians or citizens.

Perhaps we should take a breather from the eager anticipation of the future possibilities and of your wonderful new technologies and consider what is at stake.

The unrestrained exploitation of fourth-generation computer hardware and powerful new software could, I believe, undermine the very delicate balance which is the foundation of a free society. This is the balance between individual freedom and diversity on one hand--and central control and efficiency on the other.

In his book, **Dossier Society,** Kenneth C. Landon expresses it this way:

> Power is limited by segmenting authority, segregating information flows, creating multiple checkpoints, and encouraging lengthy and slow deliberation. These practical principles of political democracy are very much at odds with the virtues of contemporary information technology."

In Canada we have, as a matter of information policy, put limits on the ability of government institutions to match or link personalized data. The **Privacy Act** enshrines this policy in the following terms:

> "Personal information under the control of a government institution shall not, without the consent of the individual to whom it relates, be used by the institution except .... for the purpose for which the information was obtained or compiled by the institution or for a use consistent with that purpose...".

The requirement is not easy to fulfill when the data in question is to be acquired from or provided to another department. The reason is the **Privacy Act** places legal and procedural obligation on departments which wish to share personal information with Stats Canada.

Even in situations where administrative data can be legally acquired for statistical purposes, I am uncomfortable. The primary reason is my sense that the accuracy, completeness and clarity -- what can be termed the "quality" -- of such data, is suspect.

Despite the growing importance of data quality in large record systems, it is my sense that methodologies for examining record quality have not been established and few systematic research efforts have attempted to establish empirical levels of data quality.

To my knowledge no such systematic research efforts have taken place in Canada. In the U.S., research into the quality of FBI criminal history records between 1979 and 1982 found only 25.7 per cent of records were complete, accurate and unambiguous.

The FBI also examined some 453 warrants, issued in Mobile, Alabama as part of its 1984 data quality assurance program. Of those, 338 listed height as 7 feet, 11 inches, weight as 499 pounds and hair as "XXX". Since physical descriptions play such a significant role in law enforcement, these findings are in no way amusing. I use this example of unacceptable data quality because it comes from a records system whose very nature should dictate more rigorous quality control than in most systems of administrative records.

The quality of personal information records held by government will become an area of special interest to me as I proceed to audit government institutions' compliance with the **Privacy Act.**

Subsection 6 (2) of the Act states:

> "A government institution shall take all reasonable steps to ensure that personal information that is used for an administrative purpose by the institution is as accurate, up-to-date and complete as possible."

What steps are reasonable will, of course, vary depending on class of record and purpose of use; but one thing is clear: it is a breach of the Act to have no mechanisms to ensure data quality.

It it also clear that, from our observations, the use of administrative records for statistical purposes implies a willingness to rely on sometimes very questionable data. I noted that Gerry Gates referred to receiving IRS information which, he said, was "far from perfect".

I am also concerned that, should nominative administrative data become a favoured choice for statistical research, we may lose the ability to ensure that access and correction rights are respected. The **Privacy Act** gives individuals the right to see their records and request corrections. The exercise of these rights presupposes that the individual can determine where his or her records are held and that the department can find all copies if corrections are necessary.

The more personal information is shared among departments, the more difficult, if not impossible, this task becomes. To the extent that it becomes difficult or impossible, individual privacy rights are correspondingly diminished.

While you may not agree with all my suspicions, perhaps I could ask you to consider at least the following questions. The answers are fundamental to sensitive data collection and effective privacy protection, wherever you live and work. The principles are at the base of all international privacy protection codes and you should be able to answer "yes" to each of these questions.

> Question 1. Do you have an inventory of all your administrative data bases containing personalized information and how you are using them?
>
> Question 2. Do your administrative data subjects know how their information is being used? And do they consent to this use?
>
> Question 3. Have you taken all reasonable steps to avoid disclosure? Have you considered encrypting all your administrative data to ensure identification would be impossible if the data fell into the wrong hands?

Question 4. Can you be sure of the accuracy of the data and can you correct it if the subject can prove it is wrong?

Question 5. Can you provide the subject access to the data?

Finally, do you have consistent, widely communicated formal policies on the use of personal identifiers in administrative data?

Widespread knowledge and application of data protection principles is the only way for statistical agencies to avoid handing their citizens the kind of surprise the Swedes received last year. And my guess is that as North Americans with a less benign view of government, our reaction would be a lot tougher.

Well, this is the view from the other side of the fence. This is one fence I won't sit on. Some of you may find this viewpoint obstructionist, even paranoid. But it is sufficiently widely shared that the U.S. Congress is about to pass a bill to reverse the trend to linking the vast administrative data bases that were making a mockery of the American Privacy Act.

The bill, known as the Computer Matching and Privacy Protection Act of 1986, will require agencies exchanging data to enter written agreements. These agreements must set out the purpose, justification and legal authority for the match, describe the matching program in detail, notify the individuals whose records are being matched, verify the accuracy of the new records and ensure their security. The bill would also establish data integrity boards to oversee government computer matching programs.

Our own government has recognized data linkage as a real threat to individual privacy, that citizens should not be stripped to their souls by technology and statistical methods. It committed itself only last month to a policy spelling out the ground rules. One of these rules will require agencies to notify me 60 days in advance of a proposed match so that I may recommend either that it not take place or tell the public that it will take place. Another initiative, by the way, one that may complicate your lives, is the government's commitment to restrict the use of the SIN as a de facto federal government identifier.

I don't want you to read my comments here today as a an interdiction against any form of data linkage. The business of government is ever more a big business. And there are competing goods which Parliament must decide. But I am pleased that Parliament has recognized that their representative, and the public's privacy ombudsman has a role to play.

In the final analysis, statisticians hold the keys to their own fate. Reasonable access to direct and indirect sources of personal data depends on continued sensitivity to and respect for privacy protection principles, for values more important than mere efficiency. I believe you share that commitment.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

REMARKS AT PANEL DISCUSSION

T.B. JABINE[1]

My presentation will be in two parts. The first part might be described as "meteorological". It consists of some brief remarks about changes in the climate in the United States since 1970 that have affected statistical and other uses of administrative records. In the second part, I will present and discuss some results of survey questions on public attitudes about statistical and other uses of administrative records in the United States. These results have just become available and should begin to fill some of the gaps in our knowledge of what the public thinks about the kinds of activities we have been discussing at this symposium.

The Census Bureau and other "old-line" U.S. statistical agencies have always felt strongly about preserving the confidentiality of the individual data on which their statistics are based, whether the data are collected in surveys or taken from administrative sources. In the 1970's however, confidentiality and data access questions assumed new importance for all statistical agencies and for custodians and other users of administrative data, as the result of a series of landmark events:

- Issuance, in 1973, of the report of the Advisory Committee on Automated Personal Data Systems to the U.S. Secretary of Health, Education and Welfare. The report, **Records, computers and the Rights of Citizens,** recommended a series of legislative and administrative steps to establish fair information practices of personal record systems maintained by the U.S. Government.
- Passage, in 1974, of the Privacy Act, which enacted into law many of the HEW Secretary's Advisory Committee's recommendations. The act is similar, in many respects, to more recent Canadian legislation.
- The studies undertaken and reports issued by the Privacy Protection Study Commission during the period 1975-1977.
- Passage of the Tax Reform Act of 1976 (not to be confused with the landmark Tax Reform Act of 1986). The 1976 Act restricted non-tax uses of Internal Revenue Service (IRS) data by other agencies to a small number of purposes explicitly mentioned in the Act.

These events created a new environment for statistical uses of administrative records. Barriers were created to some important uses: for example, release of microdata files based on the Social Security Administration's Continuous Work History Sample was judged to be no longer possible under the provisions of the 1976 Tax Reform Act (Buckler and Smith, 1980). As required by the Privacy Act, agencies obtaining data from idividuals, whether for administrative or statistical uses or both, began informing people more fully about the conditions under which they were being asked to supply data and the uses that would be made of their data. Agencies gradually became more aware of the possibility

[1] Thomas B. Jabine, Statistical Consultant, 3231 Worthington Street North West, Washington, D.C. 20015, U.S.A.

that unintended disclosure of individually identifiable data might occur as a result of the release of aggregate or microdata for statistical uses. Consequently, they reviewed their data dissemination policies and generally cut back on the amount of detail that could be included in tabulations and microdata files released to the general public.

These trends have carried over into the 1980's. Some new factors have emerged. There has been strong opposition to population censuses and some types of surveys in several Western European countries, and in some instances, a drop in the level and quality of response. The development of effective record linkage methodology, using powerful computers, has increased the possibility that an "attacker" determined to identify one or more persons whose records were included in a public-use microdata file could, in fact, do so (Jabine and Scheuren, 1986).

Strangely enough in this environment, there has been increased use of record linkage for compliance purposes. The U.S. Congress has required the IRS to make tax data available to state agencies to help identify welfare recipients who are ineligible and persons who are deliquent in making child support or loan payments. Statistical agencies, however, are exercising great caution about undertaking activities that might appear to be less threatening to the public. Standards for the release of microdata files are becoming more stringent and there is considerable reluctance to undertake new record linkages for statistical purposes, even where precendents exist.

Official statisticians justify these policies as essential to protect their primary asset: The trust and confidence of the public, which leads the great majority of persons to cooperate in major censuses and surveys. Most of us will agree that public attitudes and perceptions about data collection, sharing and dissemination play an important role in determining what can be done and it follows that it is important to learn as much as we can about these attitudes and how they change over time.

We recently had a chance to collect some data on public attitudes about interagency data sharing by including a set of questions on this subject in a taxpayer Attitudes Survey conducted in the Summer of 1987 by Louis Harris and Associates for the IRS. Before describing the results, I have two general comments about the collection of survey data on public attitudes. First, in contrast to some other types of survey data discussed in this symposium, the quality of data on attitudes cannot be checked by comparing them with administrative data. Second, the survey variables are largely defined by the specific wording of questions, the manner in which alternative responses are presented to respondents and the broad context in which the questions are placed. For this reason the handout (see attachment) shows the results in conjunction with the relevant section of the survey questionnaire. I will be describing only the results of questions 97 to 99, which dealt with interagency data sharing.

The target population for this survey consisted of people who normally file federal income tax returns. For couples filing joint returns, only the more knowledgeable of the two joint filers was included. Data were collected in face-to-face interviews. In developing the set of questions on data sharing, we started with the assumption that very few people had given much thought to the question of interagency data sharing or were aware of current practices. Therefore, we approached the subject in three stages. First, we asked a series of question (Q.97, a to f) about underlying factors that might determine a respondent's attitude toward data sharing. Then we asked a single question (Q. 98a) to determine the respondent's attitude toward interagency data sharing in the abstract. Respondents were also given an opportunity (Q. 98b) to explain why they held that particular view. Finally, the respondent was asked for his or her views on the transfer of IRS data to selected agencies for specified purposes (Q. 99, a to d). Four such examples were included, two involving transfer for statistical purposes (Q. 99, a and d) and two involving transfer for administrative purposes (Q. 99 b and c). To average out response order effects, a random procedure was used to vary the order in which the examples were presented to respondents.

Looking at the results, we can see that substantial majorities wanted to know which agencies have information about them (Q.97a) and why these agencies want the information (Q.97b). Smaller majorities agreed that sharing data would reduce response burden on the public (Q.97d) and the cost to the government of getting the information it needs (Q. 97e). On the subject of trusting some agencies more than others (Q. 97, c and f), the results were somewhat inconclusive and were affected by the manner in which the question was put.

The results of the general question on attitudes toward data sharing (Q.98a) were consistent with the results of the preceding questions about factors that might influence attitudes toward data sharing. Leaving out those who were undecided or did not answer, about the same proportion favored and opposed data sharing (38 percent vs. 41 percent), but the proportion **strongly** opposed was substantially greater than the proportion strongly in favor (23 percent vs. 14 percent).

In spite of these somewhat negative views on the overall question about data sharing, we found that when we asked about the transfer of IRS data to specific agencies for specific purposes, substantially more people favored these transfers than opposed them, regardless of whether the intended uses of the data were for statistical or compliance purposes. Still, in each instance about 1 in 6 persons were strongly opposed to sharing data.

Further analysis of these data, in conjunction with other survey variables, is underway. There is one other result, not too surprising, that I can share with you, namely, that people who have had unfavorable experiences with IRS or have negative views about IRS are more likely to oppose data sharing.

We think we have managed to get a fairly good snapshot or current views about interagency data sharing. However, the rapidity with which views on subjects like this can change has been demonstrated many times. One example has to do with the proposed "Australia card" mentioned by Dr. Redfern in his paper. As the result of an articulate campaign by opponents of a national identity card, public support for the proposal declined from 65 to 39 percent over a nine-month period (**Washington Post** 1987).

In summary, our results show that strong opponents of record sharing, while in the minority, outnumber strong proponents. As a statistician, I favor expansion of statistical uses of administrative data, but the survey findings suggest that we proceed with caution. We should continue to collect data on public views of interagency data sharing and related topics, and I would be the first to admit that improvements in the specific questions are possible.

Note: For references cited above, see the symposium paper by T. Jabine and F. Scheuren, "Statistical Uses of Administrative Records in the United States: Where Are We and Where Are We Going?".

ATTACHMENT

SELECTED RESULTS FROM THE 1987 TAXPAYER ATTITUDES SURVEY

Now we want to find out what you think should be done when different parts of government ask for the same information about you. For instance...IRS, Social Security and the Census Bureau all collect information about your income. They use it for different purposes ... collecting taxes, giving benefits, finding out how our nation's incomes are changing. Some people think these agencies should **share** the information; others think each one should get it separately from you.

(HAND CARD "R")

97. Using the scale on this card, please tell me how you feel about each statement by choosing any of the four points from strongly agree to strongly disagree. You may read me the letter of the statement and the number of the scale.

	Strongly Agree	Agree	Neutral No Opinion (Vol.)	Disagree	Strongly Disagree	Not Sure
a. I want to know which agencies have the information about me	(59(50-1	27-2	16-3	4-4	1-5	3-6
b. I want to know why each agency wants the information	(60(53-1	29-2	12-3	3-4	1-5	3-6
c. I trust some agencies but not others	(61(21-1	27-2	26-3	14-4	6-5	5-6
d. Sharing would reduce the burden on the public of filling forms or answering questions	(62(19-1	33-2	20-3	15-4	6-5	7-6
e. Sharing would reduce the government's cost of getting information it needs	(63(22-1	35-2	17-3	14-4	5-5	7-6
f. One agency can be trusted as such as another in this country. They are all part of the same government	(64-13-1	26-2	20-3	23-4	13-5	5-6

n = 2,003

98a. All things considered, how do you feel about different agencies sharing information about you when they want the same information?

Favor sharing strongly	(65(14-1
Favor sharing somewhat	24-2
Undecided	20-3
Oppose sharing somewhat	18-4
Oppose sharing strongly	23-5
No Answer	1

98b. Why do you feel that way?

_____ (66-67)
_____ (68-69)
_____ (70-71)

99. You have given me your general views about different agencies sharing information. Now I'd like to ask some specific questions about the IRS sharing its information with other agencies. For each of these departments and purposes, tell me how you feel about the IRS sharing information. Do you favor strongly, favor somewhat, oppose somewhat or oppose strongly the IRS sharing information with (READ EACH ITEM).

	Favor Strongly	Favor Somewhat	Oppose Somewhat	Oppose Strongly	Neutral (Vol.)	Not Sure
ROTATE —START AT "X"						
( )a. The Census Bureau for studying population trends	(72(27-1	34-2	11-3	16-4	6-5	6-6
( )b. The Department of Justice for major criminal investigations (such as drugs and organized crimes)	(73(37-1	28-2	10-3	16-4	4-5	6-6
( )c. State governments for improving state tax collection	(74(22-1	31-2	14-3	19-4	7-5	8-6
( )d. The Commerce Department for studying economic trends	(75(23-1	33-2	12-3	16-4	7-5	9-6

INTERVIEWER: INDICATE WHICH ITEM WAS STARTING POINT FOR ROTATION: ABOVE

Item a	(76(27-1
Item b	22-2
Item c	24-3
Item d	17-4
Not indicated	10

100a. How well do you think the IRS protects the confidentiality of the information you given the government for tax purposes? Do you think it does a (READ and RECORD)?

Very good job	(77(11-1	
Good job	31-2	
Just fair job	23-3	
Poor job	12-4	
Not sure	23-5	782

N = 2,003 for all questions

100b. Why do you feel that way?

_____ (79-80)
_____ (0*-10-11)
_____ (12-13)

101. Cheating can be reduced if some government agencies give information they have about persons to IRS. For each of the following agencies, please tell me how you feel about their sharing the information they have with IRS. Do you favor strongly, favor somewhat, oppose somewhat, or oppose strongly the IRS **receiving** information from (READ EACH ITEM)?

	Favor Strongly	Favor Somewhat	Neutral (Vol.)	Oppose Strongly	Oppose Somewhat	Not Sure
a. The Department of Education	(14(19-1	26-2	16-3	14-4	18-5	7-6
b. Social Security	(15(23-1	31-2	12-3	13-4	16-5	5-6
c. Welfare agencies	(16(29-1	30-2	10-3	11-4	15-5	6-6
d. Veterans Administration	(17(23-1	27-2	14-3	12-4	16-5	7-6

102a. Were you aware that the IRS collects delinquent debts such as child support payments, student and VA loans, and other debts by retaining some or all the refund due to the taxpayer?

Yes	(18-42-1
No	51-2
Not sure	7-3

102.b. Do you favor or oppose IRS helping other parts of the government to collect unpaid debts in this way?

Favor strongly	(19(28-1
Favor somewhat	25-2
Undecided	18-3
Oppose somewhat	13-4
Oppose strongly	16-5
No answer	*

102c. Why do you feel that way?

_____ (20-21)
_____ (22-23)
_____ (24-25)

* less than 0.5 %
n = 2003 for all questions.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

PRIVACY ISSUES INVOLVED IN THE EXPLOITATION OF ADMINISTRATIVE RECORDS FOR STATISTICAL PURPOSES

G. LABOSSIÈRE[1]

I want to say a few words this morning about the privacy issues involved in the exploitation of administrative records for statistical purposes from the perspective of a senior manager responsible for controlling access to and use made of administrative records in a statistical organization.

From the previous discussions at this symposium, it is quite clear that, in Canada and elsewhere, administrative records are becoming an increasingly important part of the fabric of our statistical systems. We have heard that current uses of administrative records include direct tabulation, indirect estimation, substitution for survey responses, frame construction and maintenance, and data evaluation. These uses now permeate most statistical programs, and all indications point to more intensive and broader exploitation of these records to maintain programs in the face of declining real budgets, to produce information of greater depth and breadth, as well as for respondent burden considerations. Therefore - in summary, the use of administrative records for statistical purposes can have significant benefits - more and better information, greater efficiency, cost reductions -and respondent burden decrease.

Now, information production, particularly statistical information production implies data collection. And data collection by nature, is intrusive, and is often resented by the public which is asked to supply these data. In some instances, this resentment may not be directly related to the specific information being collected, but represents an instinctive reaction to perceptions of erosion of autonomy, or violation of private space. In the absolute, an individual's right to privacy implies the ability to control the right of access to information about one's self, and whether this information can be given free circulation, limited circulation, or no circulation. At this extreme, any collection of information is an invasion of privacy.

However, it is generally and realistically accepted that the legitimate needs of society for information - to condition its policy making and to guide its decision-making -should override such an extreme concept of privacy.

Legislation such as the Canadian Statistics Act which effectively **requires** respondents to provide information is, in this sense, an articulation of such a public good or public interest. The Canadian Privacy Act also recognizes that personal rights may have to be superseded when a more broadly-based need has to be satisfied when it permits the collection of personal information by a government institution if such information relates directly to an operating program or activity of the institution.

As different from several Western European countries where data collection plans must be approved by Data Protection Agencies, in this country, the authority to create new information banks is effectively delegated to the heads of government institutions, subject to a technical review by a central clearing-house attached to Statistics Canada.

---

[1] Guy Labossière, Statistics Canada, 26th Floor, Section P, R.H. Coats Building, Ottawa, Ontario, K1A 0T6

At the outset of a new data collection activity, therefore, it is up to the agency involved to apply its institutional conscience in the area of privacy when making a decision to collect data on a new dimension of economic or social activity - whether by direct surveying, or by accessing records already collected for other purposes by third parties.

When considering privacy issues involved in data collection, from a purely self-interest point of view, a statistical organization, if it wants to survive, must look beyond minimum compliance with legal statutes, and consider the impact of its activities on the voluntary and willing cooperation of the public to data requests. To promote such cooperation, and needed climate of trust, it is imperative that a number of conditions be clearly defined and given visibility, including:

- the legal authority to collect the data
- an explanation and a justification to the respondent of the public need to be served by the data, and of the uses to which it will be put - in order to get his cooperation, if not his commitment
- a commitment that the information supplied will not, under any circumstances, be provided to any other party in a manner which could permit connecting up an information release with a specific data provider
- a demonstrated responsiveness to respondent complaints and criticisms related to format, detail, duplication, or burden issues.

In the case of administrative records, there is a specific legal basis in the Statistics Act for accessing this type of records, and the Privacy Act does permit the transfer of personal information from the control of one organization to another one, for research or statistical purposes - when the purpose to be served requires information in identifiable format.

However, when considering privacy issues related to data acquisitions from adminstrative record sources, one has to take into account that this involves obtaining information from another institution that collected it for its own purposes, and using these records for secondary purposes without the knowledge of the individuals to whom the information relates.

To recognize this particular characteristic of administrative records, and to offset to some degree the related privacy concerns, which can arise, over and above satisfaction of the legal requirements involved, at Statistics Canada we are now developing formal agreements between ourselves and the agencies from which we obtain administrative records, which

- will specify the nature and scope of the information wanted
- will affirm formally that the information requested will be used only for statistical purposes
- will clearly identify protection measures, limitations on uses, and designate officials to monitor and control shipment dates and methodology, utilization, and related issues.

Additionally, where statistical or research uses of administrative data involve the linking of administrative records with other administrative records or with survey data relating to the same respondent or unit of observation, further privacy concerns, real or perceived, can arise related to the creation of data banks which pool information about individuals from a variety of sources, and also to the linkage of records initially collected for totally unrelated purposes. After intensive internal review and wide consultation about

the propriety of carrying out such linkages, Statistics Canada has adopted a policy which permits record linkages, but, only when the potential public good of the resulting **statistical information** is judged to clearly outweigh the risks of invasion of privacy. All record linkage proposals within Statistics Canada must satisfy all the following conditions:

- the purpose of the record linkage activity must be of a statistical or research nature and must be consistent with the mandate of Statistics Canada as described in the Statistics Act; and
- the products of the record linkage activity will be released only in accordance with the confidentiality provisions of the Statistics Act and with any applicable requirements of the Privacy Act; and
- the record linkage activity must have demonstrable cost or respondent burden savings over other alternatives, or is the only feasible option; and
- the record linkage activity will not be used for purposes that can be detrimental to the individuals involved, and the benefits to be derived from such a linkage must be clearly in the public interest
- the record linkage activity has to be judged as not jeopardizing the future conduct of Statistics Canada's programs; and
- the linkage must satisfy a prescribed review and approval process.

The review and approval process involves the submission of formal documented proposals for record linkages by directors of Statistics Canada's subject-matter divisions to a formal department committee. The recommendations of this committee are forwarded to the Chief Statistician who refers for ministerial approval all recommendations he supports.

This high level formal review and approval process is indicative of the importance attached to ensuring that linkage proposals are well documented, are considered in relation to practical, visible criteria within a formal process on a case by case basis, and that the final judgment on the balance between the public good benefits of the application and the potential invasion of privacy is made by the Minister, as a proxy for the general public.

In summary, while Statistics Canada is endeavouring to improve several dimensions of its information production role by the exploitation of administrative records, it has also recognized the particular nature of administrative records from the privacy viewpoint and is implementing measures going beyond compliance with the strict legal requirements - to achieve what it believes to be an appropriate balance in the conflicting tensions between privacy rights and official data collection, particularly where such data collection is done on a mandatory basis. As you have seen, our current position on the use of administrative records for statistical purposes reflects a great deal of care and caution, even though there is not at this time major and broadly based public hostility to such use. In this area of priorities in conflict, we believe that it is of prime importance to be completely above board with the public, and to have clear positions on information needs, information sources, uses served, and disclosure practices to dispel misconceptions or false perceptions, and hopefully to deal with potential concerns on a preventive basis.

Statistics Canada Symposium
on Statistical Uses of
Administrative Data
November 1987

PANEL DISCUSSION:

PRIVACY AND CONFIDENTIALITY CONSIDERATIONS IN THE USE OF ADMINISTRATIVE RECORDS FOR STATISTICAL PURPOSES?

J.M. LEYES[1]

The **Statistics Act** gives Statistics Canada the specific right to access administrative records, and the general responsibility to respect and protect the personal or confidential information contained in its data holdings, including administrative records.

Since the responsibility is general, Statistics Canada has developed general procedures and methodologies for protecting the private and confidential data to which it has access. Furthermore, it has developed these policies and procedures within the constraints of its operational objectives.

Clearly, administrative records have contrasting sides. On the one hand, they offer an opportunity for deriving statistics. On the other hand, they represent an opportunity for statisticians to invade the private lives of individuals in massive ways. Is it possible that the general procedures and/or methodologies used by Statistics Canada to protect privacy and confidentiality are inadequate in the domain of administrative records? Stated differently, is there an obligation to respect these incursions in ways that transcend the general procedures and methodologies that are in place for respecting privacy and confidentiality?

Prior to the creation of the Office of the Privacy Commissioner in Canada, Statistics Canada had considerable latitude in defining privacy and confidentiality. This morning in his remarks, Mr. Grace, the Privacy Commissioner, made comments and posed questions that indicate that Statistics Canada's policies, procedures and knowledge of the uses of administrative records are matters that fall within the Privacy Commissioner's domain of interest. In other words, although we, as statisticians, may pursue our work with utmost honesty and altruism in the twin domains of privacy and confidentiality, we now have someone looking over our shoulder.

Perhaps we could ask ourselves two specific quesitons about how we handle administrative records, questions that were implied but not stated directly by Mr. Grace:

> Has Statistics Canada considered the introduction of a formal policy on the use, retention and /or encryption of personal identifiers?
>
> Has Statistics Canada considered the possibility of encrypting all of the data contained in administrative records?

There is another question that it behooves us, as statisticians, to address as well:

> Has Statistics Canada considered the conduct of a regular survey and/or the use of focus groups to determine the attitudes of Canadians towards the use of administrative records for statistical purposes?

It was particularly interesting to learn from Mr. Tom Jabine in his comments this morning that the Internal Revenue Service regularly obtains attitudinal information. As

1 John M. Leyes, Statistics Canada, Ottawa, Ontario, Canada.

we all know, the IRS has a right to tax records as the income tax collection agency in the United States. Statistics Canada, by contrast, is solely a statistical agency.

To the best of my knowledge, no formal or systematic effort has been made by Statistics Canada to obtain attitudinal information on the access to and use of administrative records for statistical uses. Although we collect and disseminate data on many topics, systematic and consistent information are not available on the attitudes of Canadians regarding the statistical uses of administrative records.

In fact, it is my belief that we generally operate on the basis of faith:

- We have faith in our ability to protect the privacy and confidentiality of our respondents.

- We have faith that our respondents know that we are protecting their confidentiality.

When we collect data directly, we can maintain that our respondents knowingly provide the information. With administrative records, the scenario is different. We may be obtaining information that is very accurate, but information that was not freely given. Some Canadian residents may have provided information under the assumption that the personal information would not be used for any other purpose, including statistical applications.

While it is clear that we have an enormous opportunity to exploit the potential of administrative records for statistical purposes, we have, at the same time, the obligation to respect the personal and confidential information at our disposal, perhaps in ways that may be deemed inefficient and unnecessarily expensive. Tom Jabine stated this thought very concisely: We should proceed with caution in spite of the statistical potential that is offered by administrative records.

In the final analysis, John Grace has issued us a challenge: Statistics Canada must obtain data legally, that administrative agencies should indicate on their forms that the data will be available to Statistics Canada for statistical purposes.

What is particularly interesting to me is that strict conformance with the **Privacy Act** may inhibit or even deny Statistics Canada its current right of access to administrative records under the **Statistics Act.** For, if we do not adhere to the standard of privacy and confidentiality that lies in the hearts and minds of Canadian residents, and in spite of the papers presented this morning and in other sessions on the glowing potential of administrative records, Statistics Canada may lose access to those records.

In conclusion, given the uncertainty regarding access to administrative records and the additional costs that may be required to protect the privacy and confidentiality of Canadians, is it possible that this price is too high? Is it possible that a substantial portion of the statistical system could be placed in jeopardy through reliance on administrative records for statistical applications?

CLOSING REMARKS

CLOSING REMARKS

GORDON BRACKSTONE

"I think this has been a unique occasion in terms of bringing together people, from a wide variety of disciplines and a wide variety of application areas, in a single conference focused on the Statistical Uses of Administrative Data. We have heard about applications of administrative data in the areas of Health and Education, Socio-economic data, Business data and so on. We have heard about the methods and techniques being used to make use of administrative data including record linkage, but other methods also, and we have had discussions of the public perception issues, the issues of privacy and confidentiality. From my point of view, and from what I have heard from lots of other people, it has been a very useful opportunity to interact, to hear what is going on in other fields, and to make contacts which we hope will be productive in the future."

SURVEY METHODOLOGY

A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics.

EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, 4th Floor, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Two nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is $30.00 per year in Canada, $35.00 per year for other countries (payment to be made in Canadian funds or equivalent). Subscription order should be sent to: Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price of US $16.00 ($20.00 Can.) is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada. Please subscribe through your organization.