C.3

Catalogue no. 11-522-XPE

SYMPOSIUM 97

New Directions in Surveys and Censuses

PROCEEDINGS

Statistics Canada Statistique Canada

Canada

Data in many forms

Statistics Canada disseminates data in a variety of forms. In addition to publications, both standard and special tabulations are offered. Data are available on the Internet, compact disc, diskette, computer printouts, microfiche and microfilm, and magnetic tape. Maps and other geographic reference materials are available for some types of data. Direct online access to aggregated information is possible through CANSIM, Statistics Canada's machine-readable database and retrieval system.

How to obtain more information

Inquiries about this publication and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-8615) or to the Statistics Canada Regional Reference Centre in:

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

You can also visit our World Wide Web site: http://www.statcan.ca

Toll-free access is provided **for all users who reside outside the local dialling area** of any of the Regional Reference Centres.

National enquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Order-only line (Canada and United States)	**1 800 267-6677**

Ordering/Subscription information

All prices exclude sales tax

Catalogue no. 11-522-XPE, is published in a **paper version** for $80.00 in Canada. Outside Canada the cost is US$80.00.

Please send orders to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, K1A 0T6 or by dialling **(613) 951-7277** or **1 800 700-1033**, by fax **(613) 951-1584** or **1 800 889-9734** or by Internet: order@statcan.ca. For change of address, please provide both old and new addresses. Statistics Canada publications may also be purchased from authorized agents, bookstores and local Statistics Canada offices.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact your nearest Statistics Canada Regional Reference Centre.

Statistics Canada
Methodology Branch

SYMPOSIUM 97

New Directions in Surveys and Censuses

PROCEEDINGS

Published by authority of the Minister responsible for Statistics Canada

December 1998

Catalogue no. 11-522-XPE

Frequency: Occasional

ISBN 0-660-17408-1

Ottawa

La version française de cette publication est disponible sur demande (n° 11-522-XPF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Canadian Cataloguing in Publication Data

Symposium 97, New Directions in Surveys and Censuses (1997 : Ottawa, Ont.)
Symposium 97, New Directions in Surveys and Censuses : proceedings

Issued also in French under title:
Symposium 97, Nouvelles orientations pour les enquêtes et les recensements
ISBN 0-660-17062-0
CS11-522-XPE

1. Statistics -- Congresses. 2. Statistics -- Methodology -- Congresses.
3. Statistical services -- Congresses. I. Statistique Canada.
Methodology Branch. II. Title.

HA12S92 001.4'22 C97-988022-X

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences - Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.

∞

PREFACE

Symposium 97 was the fourteenth in the series of international symposia on methodological issues sponsored by Statistics Canada. The 1997 symposium was broad in scope, with a general theme and both invited and contributed presentations running in parallel sessions. It was held at the Palais des congrès in Hull from November 5 to 7 and was attended by approximately 500 people from 14 countries. A total of 75 papers were presented. The papers submitted by the authors, reformatted and translated, where applicable, are reproduced in this proceedings volume.

The organizers of Symposium 97 would like to acknowledge the contributions of the many people, too many to mention individually, who helped make it a success. Over thirty people volunteered to help in the preparation and running of the symposium, and an equal number verified the translation of papers submitted for this volume. Naturally, they would also like to thank the presenters and authors for their presentations and for putting them in written form. Finally, they would like to thank Christine Larabie, who was assisted by Denise Blair, Pauline Guenette and Claudette Marleau, for processing this manuscript.

The Symposium 97 Organizing Committee

Johanne Denis
Marc Hamel
Dave Dolson
Diane Stukel
Jack Gambino
Kathyrn Williams

STATISTICS CANADA SYMPOSIUM SERIES

1984 -	Analysis of Survey Data
1985 -	Small Area Statistics
1986 -	Missing Data in Surveys
1987 -	Statistical Uses of Administrative Data
1988 -	The Impact of High Technology on Survey Taking
1989 -	Analysis of Data in Time
1990 -	Measurement and Improvement of Data Quality
1991 -	Spatial Issues in Statistics
1992 -	Design and Analysis of Longitudinal Surveys
1993 -	International Conference on Establishment Surveys
1994 -	Re-engineering for Statistical Agencies
1995 -	From Data to Information: Methods and systems
1996 -	Nonsampling Errors
1997 -	New Directions in Surveys and Censuses

STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES
PROCEEDINGS ORDERING INFORMATION

Use the order form on this page to order additional copies of the proceedings of Symposium 97: New Directions in Surveys and Censuses. You may also order proceedings from previous Symposia. Return the completed form to:

SYMPOSIUM 97 PROCEEDINGS
STATISTICS CANADA
FINANCIAL OPERATIONS DIVISION
R.H. COATS BUILDING, 3rd FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
K1A 0T6
CANADA

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada" - Indicate on cheque or money order: Symposium 97 - Proceedings Canada).

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH	_____	@ $10
1987 -	Les utilisations statistiques des données administratives - FRENCH	_____	@ $10
1987 -	SET OF 1 ENGLISH AND 1 FRENCH	_____	@ $12 PER SET
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL	_____	@ $10
1989 -	Analysis of Data in Time - BILINGUAL	_____	@ $15
1990 -	Measurement and Improvement of Data Quality - ENGLISH	_____	@ $18
1990 -	Mesure et amélioration de la qualité des données - FRENCH	_____	@ $18
1991 -	Spatial Issues in Statistics - ENGLISH	_____	@ $20
1991 -	Questions spatiales liées aux statistiques - FRENCH	_____	@ $20
1992 -	Design and Analysis of Longitudinal Surveys - ENGLISH	_____	@ $22
1992 -	Conception et analyse des enquêtes longitudinales - FRENCH	_____	@ $22
1993 -	International Conference on Establishment Surveys - ENGLISH (available in English only, published in U.S.A.)	_____	@ $58
1994 -	Re-engineering for Statistical Agencies - ENGLISH	_____	@ $53
1994 -	Restructuration pour les organismes de statistique - FRENCH	_____	@ $53
1995 -	From Data to Information - Methods and Systems - ENGLISH	_____	@ $53
1995 -	Des données à l'information - Méthodes et systèmes - FRENCH	_____	@ $53
1996 -	Nonsampling Errors - ENGLISH	_____	@ $55
1996 -	Erreurs non dues à l'échantillonnage - FRENCH	_____	@ $55
1997 -	New Directions in Surveys and Censuses - ENGLISH	_____	@ $80
1997 -	Nouvelles orientations pour les enquêtes et les recensements - FRENCH	_____	@ $80

PLEASE ADD THE GOODS AND SERVICES TAX (7%)
(Residents of Canada only) $ _____

TOTAL AMOUNT OF ORDER $ _____

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER

NAME ______________________________

ADDRESS ______________________________

____

CITY ____________ PROV/STATE ____________ COUNTRY ____________

POSTAL CODE ____________ TELEPHONE (______)____________ FAX ____________

For more information please contact John Kovar. Telephone (613) 951-8615. Facsimile (613) 951-1462.

NEW DIRECTIONS IN SURVEYS AND CENSUSES

TABLE OF CONTENTS[1]

[1] In cases of joint authorship, the name of the presenter is shown **boldface**.

SESSION I-4: Coverage Measurement in Censuses
Chairperson: B. Laroche, Statistics Canada

SESSION I-5: Estimation and Variance Meet the 21st Century: Computer Intensive Methods and Resampling
Chairperson: C.-E. Särndal, Statistics Canada

SESSION I-6: Data Warehousing and Internet Dissemination
Chairperson: M. Podehl, Statistics Canada

SESSION I-7: Geography and Surveys
Chairperson: B. Ruth, Statistics Canada

SESSION I-8: Advances in Imputation
Chairperson: J. Kovar, Statistics Canada

SESSION I-9: Data Quality and Improving Data Processing
Chairperson: K. Bennett, Statistics Canada

SESSION I-10: Pushing the Frontiers of Design and Estimation: Data from Multiple Sources
Chairperson: J.N.K. Rao, Carleton University, Canada

SESSION I-11: New Directions in Questionnaire Development, Testing and Design
Chairperson: L. Swain, Statistics Canada

OTHER INVITED PRESENTATIONS

PANEL DISCUSSION: Approaches to Innovation in Statistical Agencies

CONTRIBUTED PAPERS

SESSION C-1: Imaging, Integration and Automation in Data Processing

Chairperson: M. Turner, Statistics Canada

SESSION C-2: Topics in Estimation and Variance

Chairperson: D. Binder, Statistics Canada

SESSION C-3: International Experiences in Sample Design

Chairperson: M.P. Singh, Statistics Canada

SESSION C-4: Dealing with Non-Response
Chairperson: S. Michaud, Statistics Canada

SESSION C-5: Handling Census Information
Chairperson: G. Hole, Statistics Canada

SESSION C-6: Sample Design and Sample Selection Methodologies
Chairperson: N. Laniel, Statistics Canada

SESSION C-7: New Data Collection Technology Adapting to a Changing World
Chairperson: B. Williams, Statistics Canada

SESSION C-8: Tackling Response Burden
Chairperson: J.-F. Gosselin, Statistics Canada

SESSION C-9: Data Analysis: Where We're at and What's New?
Chairperson: J.-M. Berthelot, Statistics Canada

OPENING REMARKS

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

OPENING REMARKS

G.J. Brackstone[1]

On behalf of Statistics Canada, welcome to this the 14th Symposium in a series that goes back to 1984. For those of you from outside the country welcome to Canada, for those of you from out of town welcome to Canada's National Capital Region, and for those of you from Ottawa, welcome to Hull. This is, in fact, only the second of our 14 symposia to be held outside Ottawa, and the first to be held in Hull.

This looks like it might be our biggest Symposium for many years. We have over 500 registrants. While that includes a large home crowd contingent from Statistics Canada, we have registrants from 14 different countries on 5 continents. From within Canada we are pleased to see strong representation from several federal departments (particularly Revenue Canada) and from le Bureau de la Statistique du Québec.

Over the years, our symposia have covered a wide range of topics related to survey-taking and the methodological work of statistical agencies. You will probably have seen the list of topics in the Symposium brochure. They range from small area statistics to the impact of technology to Re-engineering for statistical agencies. Some of these topics have been narrowly focused and more technical, while others have addressed broad issues from statistical, operational and managerial angles. Recently we have attempted to alternate these two types of topic. Last year we addressed the narrower topic of Non-sampling errors; this year we address the very broad topic of Future directions in surveys and censuses; next year our subject will be Longitudinal analysis for complex surveys.

So this year's topic is a very broad one. I suppose the future directions of surveys and censuses are something we would all like to be able to predict, to prepare for, and maybe to influence. But, as they say, forecasting is difficult, especially of the future. Yet at this time the changes that are happening in the world that we measure through our censuses and surveys are so profound and fast that if we do not keep looking ahead and adapting we will soon be behind.

Let me mention three evident trends in our environment that are having an important impact on the way we design and carry out censuses and surveys.

First, the topics on which information is demanded, the topics that our surveys are required to measure, are becoming more numerous and complex. They are extending into more sensitive areas that were once considered off-limits for surveys, and they are requiring data that goes beyond the purely descriptive to data that provides insights into the processes and interactions that cause changes to happen. This requires us to think anew about sources of data, about overall survey design, and about collection instruments.

Secondly, this expansion of requirements is taking place in a social climate which, in many countries, is putting more emphasis on privacy concerns which, in turn, is causing more respondents to question survey purposes, to seek clear confidentiality assurances, and to understand the intended data uses, as a basis for deciding whether to cooperate. This requires us to concentrate even more on respondent relations, on explaining the purpose and value of surveys, and, inevitably, on dealing with nonresponse.

This heightened concern about privacy is no doubt driven in part by unease brought on by the third trend, the amazing technological advances of recent years, especially in the domain of electronic communication. For survey and census takers, technology is both a threat and an opportunity. We are already well aware of the opportunities that it offers.

For example. we have seen how technology is revolutionizing data collection and capture, with computer assisted collection, in some form or other, now the norm. We can expect this rapid revolution to continue as the Internet becomes more pervasive, as voice and handwriting recognition become more cost-effective, and as the potential for businesses to provide data directly from their internal computer records is realised.

But, in the area of data collection and capture, technology presents some new difficulties too. For example, modern telecommunications technology is making the use of the telephone for household surveys much more complex. The old assumption that each household has one telephone line, and is accessible through it, becomes more and more archaic. Instead of worrying only about the small proportion of households without telephones, we have to consider answering machines, call blocking, FAX lines, data lines, cellular phones, pagers and so on.

A second example of technology influence: estimation and analysis. We are now able to utilize computer-intensive methods, such as resampling methods for variance estimation or complex analytic methods, to an extent that would have been impractical a decade ago. But there is a danger here too. The danger is the substitution of computer power for intelligence. The broader range of methods made more easily available to more people by greater computing power heightens the need for statisticians to make intelligent decisions about the applicability of different methods to particular situations.

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

As a third example, consider the changes occurring in dissemination. Today the results of our surveys and censuses are available in electronic form to all who want to receive them that way. Users can browse databases looking for what they need, and if it isn't already published they have the option of requesting some additional retrievals. This is a far cry from our complete dependence on paper products just a few years ago.

One result of being in the information age is that the sources of data are increasing rapidly. Our traditional datasets from surveys and censuses are becoming a smaller proportion of the total information barrage that hits us every day. In this environment the role of surveys and censuses themselves may be changing as more and more data become available through the automation of administrative and commercial processes. Of course, statisticians have used administrative data for years, but as technology makes this kind of data more pervasive and more practical to use, we have to rethink how traditional statistical surveys interface with these other data sources.

There are some other implications of the technology revolution and the information age that we have to keep in mind. Let me mention two.

As respondents become more and more accustomed to dealing with businesses and governments electronically, this becomes the expected mode of doing business. Survey organizations that do not offer these modes of response will be seen as old-fashioned, inefficient, and be left behind. But of course, not all respondents are ready for the latest technology, so flexibility in providing alternative response mechanisms is needed.

Secondly, the public is becoming used to faster transactions and faster service. They will expect faster turnaround of survey and census results. It will become harder to justify why surveys or censuses still take many months, or even longer, to produce results in an age of instant communication and extensive computing power. This could become one of our key challenges.

To summarize, the increasing demand for new and more complex statistical data, the rising concern for privacy, and the opportunities and threats presented by rapid technology advances, combine to make this a very appropriate time to be considering the future of censuses and surveys - without even invoking that favourite excuse for prophesy, the new millennium.

Let me turn now to our program. Jack Gambino and his team have put together a great program which, over the next three days, will take us through almost all dimensions of survey-taking.

We have sessions dealing with innovative survey design, with other sessions focused more specifically on sample design. There are several sessions dealing with data collection issues, including some focusing on the important issues of nonresponse, of the use of technology in data collection, and of questionnaire design. Processing issues are addressed in sessions on automation and on imputation, and an interesting session on Geography and Surveys will address ways in which surveys can take more advantage of the rapidly developing world of GIS.

We have a couple of sessions dealing with the special issues that arise in Censuses of Population, as well as one on the use of administrative data. Estimation, variance estimation, and analysis are all covered, and a special session concentrates on data quality issues. Finally, to complete the survey cycle we have a session on electronic dissemination.

After we have heard about the latest developments and future directions in all these aspects of survey-taking, we will end with a Panel discussion on innovation in statistical agencies. How can we ensure that new things happen when and as they should? How can we provide an environment that encourages innovation and allows new ideas to become reality?

So I hope that in this program, everyone here will find something new, something of interest, something to take home that will make a difference in your own survey practice and justify the investment of time and money that you have all made in coming here this week to support Symposium 97.

I wish you all an enjoyable and productive three days.

KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

TOWARDS MULTIPLE-MEDIA SURVEY AND CENSUS DATA: RETHINKING FUNDAMENTAL ISSUES OF DESIGN AND STATISTICAL ANALYSIS

S.E. Fienberg[1]

ABSTRACT

Sampling and survey methodology have made great strides in the past 25 years and many of the recent advances have been linked to model-based approaches to survey analysis. But the physical and political environments in which statisticians gather data are in enormous ferment and, at the same time, new forms of data are emerging as alternatives to the traditional numerical responses that survey methodologists have dutifully encoded for use in statistical analyses. This paper reviews some recent developments in survey analysis, a few of the current challenges for the design and analysis of surveys and censuses, and then speculates on future challenges.

Survey data sets of the future might well consist, either through direct collection or forms of record linkage, of combinations of traditional numbers, text, images, sound, and even symbolic summaries. New statistical methods will be needed to deal with such mixed media, and the new data and methods will raise new issues regarding the design and collection of survey data as well as their dissemination, including concerns of confidentiality and disclosure limitation. The time to begin thinking about such issues is now.

KEY WORDS: Bayesian methods; Censuses; Image analysis; Model-based sampling methods; Text data; Video-tape surveys.

1. INTRODUCTION

I am greatly honored to serve as keynote speaker for this symposium, both because of my high regard for Statistics Canada and the methodological research it has helped to foster, and because the topic of this particular symposium is so near and dear to my heart. Depending on how one decides to measure calendar time, we are just a bit over two or three years from the beginning of the next century and the next millennium (if this statement seems somewhat mysterious, see Stephen J. Gould's explanation (1997)). At any rate, I would be remiss if I did not take this occasion to reflect, at least briefly, upon the past century during which we have seen the instantiation of random sampling and survey methods in public policy and scientific thinking. I also want to speculate about some of the methodological issues that I believe will challenge us, not just for the next couple of years, but well into the next century.

It was prior to the turn of the last century, with the advocacy of Kiaer (1897) in a publication exactly 100 years ago, that sampling ideas took center stage in the discussions of large scale government data collection, especially at the meetings of the International Statistical Institute (ISI). But it was only with the 1925 report to the ISI on methods for sampling by a commission headed by Adolph Jensen (1926) (with a theoretical appendix by Arthur Bowley 1926), followed by Jerzy Neyman's classic 1934 paper read before the Royal Statistical Society, that random sampling moved foursquare into the arena of government statistical agencies (Fienberg and Tanur 1990, 1996). Thus, in a sense the real methodological history is only 60 to 70 years old. We have come a long way in the intervening time, especially over the past 25 years.

About 20 years ago, I presented my first paper on sample survey methodology at a symposium held at the University of North Carolina (Fienberg (1978)) and, in it, I proposed that the survey measurement of victimization be thought of in terms of a stochastic model for criminal victimization events over time. Such a model, I argued, should change the way we think about both the design and analysis of victimization survey data. These words were barely out of my mouth when Morris Hansen and others stood up to challenge such a perspective. I am pleased to note in the intervening 20 years, the role of statistical models in the analysis of survey data and for the collection and analysis of census data, especially in the context of adjusting for differential census undercount, *e.g.*, see Anderson and Fienberg (1988), Fienberg (1994a), and Mulry and Spencer (1993).

Model-based methods are now firmly entrenched in many government statistical agencies around the world and in topical areas ranging from economics, *e.g.*, in the analysis of gross labor flows (see Stasny (1987)), to health and nutrition, *e.g.*, in measurement error models for specific nutrient consumption (Nusser *et al.* (1996) and Chesher (1997)). Model-based approaches have rightfully taken their place beside design-based approaches and hybrids such as the model-assisted approach espoused by Särndal

[1] Stephen E. Fienberg, Maurice Falk University Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA.

(1992). For experiments embedded in surveys model-based approaches are especially useful, *e.g.*, see Fienberg and Tanur (1988) and Boruch and Terhanian (1996), but for the analysis of non-response such methods are now the "only way to go," since we need to generalize beyond the realized sample but without a basis that flows automatically from the design. Bollinger and David (1997) offer a recent example of model-based non-response analyses. The rise and acceptance of model-based approaches to sample surveys, and their movement into the mainstream of survey work along with design-based approaches, is not a new theme. For example, Smith (1994) gave an interesting account of the change during this period in his 1994 Morris Hansen lecture.

Moreover, Bayesian methods and models have begun to permeate the walls of statistical agencies despite the continuing concern voiced by some that "subjective" models and methods have no place in government activities. This has been especially the case in connection with small area estimation based on survey data, *e.g.*, see Ghosh and Rao (1994), Malec *et al.* (1997), and Stasny *et al.* (1991), and with the use of multiple imputation, *e.g.*, see Rubin (1987, 1996) and Fay (1992). There have been other methodological advances which have been influenced by developments in other fields, *e.g.*, cognitive aspects of survey design (Tanur (1992)), and yet others have been technologically driven, *e.g.*, such as CATI, CAPI, and CASIC and the use of re-sampling methodology (Fay (1984)) including multiple imputation.

In a sense, this symposium chronicles how far we have come since the 1970s, and the papers being presented here offer testimony to the remarkable innovations that have and are continuing to occur in the survey field. The titles and abstracts of many of these papers and those presented at the recent conference at the University of Nebraska, give evidence in support of my claim of a quantum shift vis a vis the role of statistical models in survey and census work and increasing attention to Bayesian approaches. There are in fact so many papers on these topics for me to reference that I have decided that diplomacy is best served by not referencing any of them and in this way I cannot insult any of my friends and professional colleagues present through the omission of references to their work.

I fervently believe in the role of random selection and probability sampling in data collection – the "gold standard" – and as the basis for proceeding with data collection into the future. Yet government statistical agencies in many countries find themselves under siege, with politicians and the public questioning fundamental aspects of the statistical science associated with surveys and censuses. In the next section, I describe some of the current challenges to sampling and survey taking. Then I turn to the new challenges that I believe we must begin to prepare ourselves for as we look to the next century, those posed by massive amounts of non-numeric data. My goal is to stimulate your thinking and to provoke. I have few pretensions that I know the real directions our collective statistical work will take many years into the future.

2. SOME CURRENT CHALLENGES

There are many current problems that offer methodological challenges. Here I focus on three that I have encountered in some of my recent activities and interests.

2.1 The U.S. Decennial Census of Population and Housing

Controversy has dogged the last two census of population and housing in the United States, largely because of proposals to correct for the differential undercount of Blacks and other minorities using the results of a post-enumeration survey, *e.g.*, see Anderson and Fienberg (1997, 1998) and Choldin (1994). And controversy threatens the forthcoming census in the year 2000 because of proposals to use sampling in the census in two ways: for non-response follow-up, and again for a post-enumeration that would provide a final correction to the count.

One of the major uses of sampling is to gain increased accuracy as a result of working more carefully with only a fraction of a relevant population and at the same time holding costs in line. This is exactly its proposed role in sampling for non-response follow-up in the 2000 census. As such it offers a modest increase in the variance of estimates of population counts in return for reduced bias and reduced census costs, *e.g.*, see the discussion in two recent reports of separate panels of the Committee on National Statistics (Edmonston and Schultze (1994), and Steffey and Bradburn (1994)). Yet despite near unanimous support from the statistical community for attempts to make census taking statistical, the Bureau of the Census has run into a roadblock on its plans from Republicans in the U.S. Congress who oppose sampling on the grounds that it is unconstitutional and that "it is subject to manipulation." Unsaid in the congressional debates on the topic is the expectation that sampling in both of its proposed forms will reduce the differential undercount and the resulting numbers, when used in the apportionment of the subsequent congress, would be of greater benefit to the Democrats than to the Republicans.

This challenge to the use of sampling in the census has potentially important implications for the use of sampling elsewhere in government. In October 1997, the House of Representatives passed a bill that would suspend plans for sampling in the census until the federal courts rule on its constitutionality, a move certain to derail the use of sampling in 2000. This bill was linked to the appropriations bill for the Department of Commerce, and the President threatened to veto it if the ban on sampling in the census is also approved by the Senate. The resulting compromise, which was enacted into law in November 1997, would in effect require the Census Bureau to test both sampling and traditional methods in the 1998 census dress rehearsal, and also allow a challenge to the constitutionality of sampling in the census to proceed through the U.S. courts. But the language of the law is strongly prejudiced against the intelligent use of sampling and statistical methods more broadly in the census context and it sets the stage for renewed political intervention into technical decisions and choices at the Census Bureau. This is not good.

Another indication of the political view of sampling which must be overcome for the long-term health of our enterprise comes from another attempt by the Republicans in the U.S. House of Representatives to require what would be a "census of tax payers" at an estimated cost of $30 million using a 14-page questionnaire on their attitudes and experiences with the Internal Revenue Service, despite the existence of a recent sample survey on the same topic whose cost was only $20,000 (Stevenson (1997)).

But something more fundamental underlies the problems leading to the U.S. differential undercount and the debates over adjustment. Virtually all surveys and censuses of population in the U.S. and elsewhere are based on the concept of a stable nuclear family living in a single household location. Evidence from the U.S. suggests that the past several decades have seen a dramatic shift in the stability of family groups, and that the census undercount problems stem in large part from the individuals with either multiple residences and family groups, or none at all, at least none that provide usable survey or census information. If this trend continues the basic approach to sampling which was institutionalized in the late 1930s and 1940s will begin to be questioned in more serious ways. And other approaches, such as those based on telephone frames, also face serious challenges in part due to the proliferation of multiple telephone numbers for individual residences, telephones dedicated to computers and FAX machines, and the rise of pagers and portable cell telephones. The challenge of finding alternative frames and approaches will remain with us into the next century.

2.2 Information for Science and Technology Policy

The Canadian government and those elsewhere in the world have recently focused considerable interest on science and technology as a basis for growth and societal welfare. One of the glaring problems has been the lack of data to support assertions and to measure outcomes. As a consequence, Statistics Canada (Advisory Committee on Science and Technology Statistics 1997) has embarked on a special project to design a science and technology information system that can be useful to inform policy in this area.

The answers to questions of interest in this field often involve the use of multiple databases, some involving surveys and others involving administrative records, where the units of measurement are often fundamentally different across the databases. The methodological problems here are thus not the standard ones involving multiple frames. Rather, they are much more complex, and involve very different uses of statistical modeling than can be found in the analysis of data from a single source.

An example of the kind of question policy makers ask in this area is: "Is there a mismatch between the education and skills of the products of Canadian schools and universities and the scientific and technological labor needs of government and industry?" How might someone go about answering such a question? They might try to use information from a longitudinal survey of university graduates, and then link that information to administrative data on firms utilizing science and technology data supplied via Revenue Canada, as well as data from other Statistics Canada surveys and censuses of establishments. Combining such data sources and reflecting both model uncertainty and data variability properly is a tremendous statistical challenge.

We see the same problems in a variety of other policy-relevant areas such as welfare and education. The creative use of administrative databases in conjunction with survey data offers many methodological challenges but also great opportunities. Scheuren (1996) describes aspects of this in the context of educational data gathered by the National Center for Educational Statistics in the United States.

2.3 The Consumer Price Index

Over the past year, considerable publicity has been given in the U.S. and elsewhere to something called The Boskin Report on the Consumer Price Index (CPI), which is based on a complex of surveys designed and carried out by the U.S. Bureau of Labor Statistics (BLS)[2]. Most of the publicity about the report focused on biases in the CPI as a measure of cost of living, amounting to what was estimated to be a 1.1% overstatement of the year-to-year change. What was not so widely discussed was the commission's support for the a new Consumer Expenditure Survey (CEX)[3] and its encouragement to BLS to consider new data collection initiatives.

Some of the circumstances surrounding the CPI and CEX illustrate a major issue facing government agencies. As important as the CPI is for government policy, it is considered by many in Congress to be a costly survey program (on the order of $25 million and up) and thus they have been unwilling to authorize expenditures to keep it fully up-to-date such as those associated with a new CEX. While it was really only with a revision completed in 1978 that probability sampling was introduced throughout the entire index (Fienberg (1994)), the elaborate structure of the CPI methodology is rooted in an earlier era. (For a brief overview of the CPI see Fienberg (1994), and for a full description of the CPI methods see Bureau of Labor Statistics (1992)).

Given the mounting costs of doing surveys like the CPI I believe that we as statisticians need to begin thinking about alternative ways to collect relevant data. In the spirit of the recommendations of the Boskin Report for new data collection initiatives, I'd like to suggest that we need to explore the usefulness of other existing databases, not necessarily under the control of federal statistical agencies.

Today, in virtually every major grocery store in the U.S. and Canada, products purchased are passed by scanners and the resulting data are recorded on company computers.

[2] *The Boskin Report* is a report to the U.S. Congress on the CPI prepared by a commission chaired by Michael Boskin, a former chair of the President's Council of Economic Advisors (1996).

[3] The CEX collects the information that produces the weights in the CPI and was last carried out in the mid-1980s.

Market research has been revolutionized by the use of such scanner data but the real revolution is only now on the horizon. Let me explain by example.

In Western Pennsylvania there are a number of grocery store chains, but only one very big one, owned by the Giant Eagle Corporation. A few years ago Giant Eagle introduced *The Advantage Card* to its customers, and offered specials and "rewards" such as a Thanksgiving turkey for those who used *The Advantage Card* when they went to the checkout counter. As a consequence, almost 80% of all Giant Eagle store purchases are made using *The Advantage Card* and thus the scanner data on these purchases can now be linked to demographic information on the purchasers and their families in one very large data base that produces about 500 MB of data a day.

I believe that scanner databases such as those of my local grocery store chain have relevance for the CPI, and need to be considered as a possible substitute for at least some of the current modes of data collection on prices and products. Because such data bases are in their infancy as research tools, what we need to do now is think in terms of exploration and assessment of them. Such an approach would require a different way of thinking than what one would find in most statistics agencies today. Only then could we begin to ask about how such data could be linked to probability samples for stores and sales locations that could not conveniently supply scanner data. My understanding is that BLS has research underway on the use of scanner data, but not necessarily in the form suggested here.

3. STATISTICAL METHODS FOR MULTIPLE-MEDIA DATA

3.1 What are Multiple-Media Survey Data?

Scientific and engineering data now come in large amounts and in new forms. New initiatives are underway to extend the methods for the analysis of purely numerical data to those that arise using other media, *e.g.*, when data elements might include symbolic logic descriptions, images, text, sound, and other media in addition to traditional numerical data. From this perspective, those involved in the collection of sample survey data, and especially the federal government statistical agencies, have not kept pace. Statistical methods for the analysis of survey data today focus sole on purely numerical data, despite the fact that some survey data already are gathered in varied forms, especially text for answers to open ended questions.

The usual statistical data base for a large scale government sample survey comes in the form of a traditional flat $n \times p$ file of numerical values, where n is the number of individuals and p is the number of variables, or some hierarchically structured version of such data reflecting families and households. In the first instance, mixed-media survey data simply means replacing numerical data elements for some column by data elements involving other media. For example, instead of using the textual or oral (sound) response for open ended questions, the survey interviewer currently "translates" or codes that information into a traditional categorical or numerical value, thus destroying potentially valuable additional information. Similarly, in some government surveys such as the U.S. National Health Examination Survey respondents give samples of blood and submit to other diagnostic tests. Test results are then extracted from the original information in the form of a numerical summary, *e.g.*, a white blood cell count, or the response to a PSA prostate cancer diagnostic test. Were the richness of the original data available, *e.g.*, in the form of something that we might refer to as a "virtual blood sample" or something else representing a recordable extract of information, new analyses might be performed at a later point in time that would take fuller advantage of the information originally collected. Were it not for the cost we might think in terms of some elaborate series of genetic tests or sequencing. The specific choice of digital representation and the substantive tools for its creation in some senses are not as important here as the idea itself.

In the sense of creating a complex digital representation, the notion of a "virtual blood sample" is not unique. After all, recorded images are also digital representations, from three dimensions into two and with pixels instead of the complete image. And other sample surveys such as the Natural Resources Inventory sponsored by the U.S. Department of Agriculture includes the taking of physical samples, in this case soil and water samples, where a different form of digital representation is used. The design of computerized data bases that will incorporate such extensive digital representations and will allow easy access and extraction of key features for statistical analysis poses new challenges for our field working at the interface of computer science. I provide some illustrations in the context of images in the following subsections.

3.2 Example: Video Surveys of Classroom Processes

Sometimes it is not the images per se, but what people are doing within them that is of interest. This is the case with the Third International Mathematics and Science Study (TIMSS) videotape classroom survey, which has focused on 8th grade mathematics and was designed to compare the teaching practices of American, German, and Japanese teachers. A total of 231 classrooms from TIMSS were randomly selected from the TIMSS country samples, 100 from Germany, 50 from Japan, and 81 from the United States. One complete lesson was videotaped in each classroom. Stigler (1996) provides considerable detail on both the questionnaire administered to the teachers and the videotape protocol. These data are also linkable, at least in principle, to the achievement testing data for classrooms in the selected schools.

The video tapes themselves are multiple-media data elements from which information is abstracted and coded. This coding has to date involved human observers and transcribers, who need to be fluent in the three different languages. The coding processes reflect the knowledge domain of those studying education (*e.g.*, the nature of the work environment, the nature of the work the students are engaged in, and the methods used by the teacher, the types of interactions between the teacher and students, *etc.*), and they extract the equivalent of functionals from the videotape for analysis by traditional methods. But since the videos have been digitized to facilitate this coding process alternative ways to working with both the images and the sound are potentially of value to those involved in the study.

Because both teachers and individual students are identifiable in the videos, only the coded information is being made available to secondary analysts, to help preserve confidentiality. But the data are being released *linked* to other information from TIMSS.

3.3 Example: Analyzing Images of the Planets

The previous example involved data from a real sample survey, but the uses of the video are not quit "high tech" in the sense of relying upon elaborate computer-based analytic manipulations. My next example moves us in this direction.

Exploration and analysis of massive data sets requires advanced methods of access to the data as well as expanded storage for the data. At Carnegie Mellon in the Department of Statistics we have recently acquired a device with one terabyte (TB) of storage to handle data sets like the scanner data referred to above, and to work with images. How do we think of the issue of access for such data sets? A major feature of such data sets is that the actual processing of the complete data can not be done very many times in any reasonable time interval, unlike the analysis of a typical survey data base where repeated processing of all of the data is commonplace. What my colleagues and others have found helpful is a focused goal to calculate a relatively small number of functions of the data for subsequent more standard sorts of processing.

Eddy and Mockus (1996) have constructed a system for interactive analysis of a very large dataset containing roughly 30 gigabytes (GB) of images, which they named *Interactive Icon Index* (I^3, or *Icecube*). The images they used in developing their system were video still pictures of the outer planets of the solar system and their moons captured by the two spacecraft, Voyager I and Voyager II. The *Icecube* system treats each image as an individual observation and it facilitates data exploration by creating an index of the images. A small copy of each image in the dataset, called an "icon" or "thumbnail", is used to represent the full image. The thumbnails are laid out on the screen of a computer, in a "contact sheet." The contact sheet is, in fact, used as the index, which is both interactive and dynamic. Figure 1 gives part of the Eddy-Mockus contact sheet for the images of Saturn. From it, one can select thumbnails, rearrange them on the contact sheet, sort them by various attributes, and so on. Figure 2 shows the full image for Saturn and its rings.

The *Icecube* system ties together four distinct databases:

1. the original database of images (stored in compressed form on CD-ROM);
2. the database of copies of the images recorded onto a video disk system (which allows near instantaneous access to each image for viewing);
3. the database of thumbnail-size icons displayed on the screen of the workstation;
4. a traditional database of numerical and alphabetic information about the images.

One can augment this latter database, in principle, with additional information computed from the images themselves.

The Icecube system as it exists currently has a number of useful features for thinking about and manipulating a database of images. Primary among those is the fact that it allows easy access to a elements of the database. It is flexible, in that thumbnails can be moved around on the contact sheet, different subsets of icons can be chosen for inspection without losing the entire index, and different attributes present in the database can be employed to sort or select data. Further, changing the arrangement of the thumbnails on the contact sheet can reveal startling patterns in the data (for instance, using an ordering that places icons that are close in linear order close together in two-dimensional space as well).

But there are also a number of desirable capabilities not currently available in the *Icecube* system. These include discarding an icon, moving an icon to an arbitrary position in the sequence, selecting all images with certain properties and selecting all images that are "similar" to a given one, and the ability to calculate functionals of the images.

3.4 Example: A Sample Survey with 3-Dimensional Images of People

Colleagues at Carnegie Mellon in collaboration with the Computerized Anthropomorphic Research and Design Laboratories at Wright-Patterson Air Force Base in Ohio, the American Society for Testing Materials (ASTM), and the Society of Automotive Engineers (SAE) are developing plans for the collection of a novel sample survey dataset. The data will consist of demographic information and three-dimensional images from three separate full-body scans taken on each of samples of approximately 10,000 individuals in the U.S. and 10,000 in The Netherlands and Italy. That is, for each of the individuals in the samples, they will obtain, in essence, three three-dimensional images of the person's body. The sizes of the scans vary but we estimate, based on several examples, that we would need about 50 MB of storage for each individual in the samples, so that the entire data set will consist of approximately one TB of data. A pretest version of the data is described in Brunsman and Files (1996). In Figure 3, I show a simple two-dimensional picture extracted from one of the images from the pretest data, but it hardly captures the richness of the recorded data. In fact, we have a graduate student developing a system which can be used to rotate the three-dimensional images using a special graphical interface.

Issues that arise from studying such data include (i) the advancement of methods for defining multidimensional size, shape and functional characteristics of humans, (ii) the developments of engineering tools and databases for designers of personnel protective equipment and clothing and (iii) the assistance in the designing, specification and testing of vehicles (airplane seating, automobiles, cockpits of military transportation). This may well be the largest and most detailed dataset ever collected for the purposes of anthropometry (the measurement of the human body).

How might one think of analyzing such data? Well we might begin by attempting to simply reuse the existing *Icecube* technology described above. But viewing a single 3-dimensional image of an individual using a standard rendering tool required 180 MB of RAM to allow storage and sorting of the hundreds of thousands of polygons which composed the surface of the individual. Second, any plausible interactive use of the system requires fairly short access times to random elements of the main database, a feature not present in the initial *Icecube* implementation. Compression of the database will reduce its size by a factor between three and four but will, we estimate, increase the access time by a similar factor. Third, any reasonable modeling activity is going

1. to be driven by fairly traditional analyses of summary statistics (functionals) computed on the objects; and
2. consist of local models which can be estimated from subsets of the data which are constructed from the interactive analyses.

Fortunately, for the proposed anthropometric database, my colleagues expect to be able to extract from the data various features and calculate some "standard" functionals of interest. For example, in an evaluation of flight suits for US Navy women, investigators found that a size 38 flight suit may be too tight in the hips and too large in the shoulders (Robinette (1995)). In this case, the minimal set of functions of the images we need to look at are shoulder width and hip breadth. But the functionals of interest today may differ from those of a later secondary analyst, and thus new systems that are developed will need greater flexibility than current ones. Some of my Carnegie Mellon colleagues are beginning such efforts.

At first blush, it still sounds that we might be able to reduce all of our statistical work to methods designed for numerical data through the choice of a sufficiently large number of functionals. But I think we need to get beyond such thinking if we are to take effective advantage of data in the form of images. Here are two ideas of what I have in mind set in the context of three-dimensional laser body scans. For both, we can think in terms of having chosen a small number of functionals, say 5, on which we might do an initial analyses. Then we could turn to

1. algorithms for creating "adjusted" images, *standardizing* or *raking* according to some common mean value for the functionals. This would lead to a new form of image-based residual analysis.
2. "typical" images to go with the means or medians of the functionals.

Both of these approaches are clearly complex and at the moment not even well-defined, especially for three-dimensional images, but they are conceptually simple and are image-like analogues to approaches we use for the analysis of numerical data.

Statisticians need to develop new methods to attack the problems associated with the analysis of massive data sets of the sort exemplified by this proposed sample survey with three-dimensional scanned images. An essential new component of the computer system required to look at individual image data elements would be the ability to continuously calculate (in the background) various functionals of the data as the interaction (in the foreground) suggests the need for them. The ability to view and interact with the individual three-dimensional images requires significant amounts of RAM in addition to that required for the background calculations.

But how do we even think of integrating tens of thousands of such records drawn from finite populations and summarizing them in digestible form? This is part of the challenge that awaits us.

4. METHODOLOGICAL RESEARCH ISSUES FOR MULTIPLE MEDIA SURVEY DATA

In more complex settings, the very notion of sampling units for mixed media becomes a concern and thus the definition of rows in our traditional $n \times p$ data array becomes an open statistical issue.

Once mixed media data become a possibility in the survey setting, new approaches to survey data open up, and images and sound also become potential data elements, especially if traditional survey data are linked with other data bases where mixed media are more the norm, *e.g.*, those stored in digital libraries or sampled from the World Wide Web.

What are the challenges of mixed media data for statistical methods? There are clearly issues of data storage and retrieval that new media raise, and survey practitioners would need to begin to think in terms of modern database tools that would allow for easy access to non-numerical information and the extraction from it of alternative numerical summaries or characterizations. By modern scientific database standards, traditional survey data sets are relatively small. Even the entire data files for the U.S. decennial census of population and housing fit on a modern PC hard drive. This will no longer be the case when images, sound, and text become integral parts of sample records. New methods to cope with such data bases will draw upon extensions of traditional statistical methodology blended with tools and methods drawn from datamining and other parts of computer science. The first class of methods will inevitably draw upon forms of data reduction, such as the computation of functionals on images, or the coding of classroom videotapes. Analyzing functionals sounds a lot like model-based inference and in the work at Carnegie Mellon and elsewhere it typically is. Thus a natural question to ask is: Is there a design-based analogue whose interpretation makes sense?

While there are well-developed methods for characterizing populations and distributions of numerical data, new statistical foundations are required to develop similar characterizations for data that include mixtures of images, text, and other types of data. For example, we lack basic notions such as appropriate definitions of mean, variance, and correlations to characterize multiple-media data samples. Further, the weighting that is typically done with survey data need some reconceptualization in the context of alternative media. And, once suitable summaries have been developed, new issues regarding their presentation, graphically and otherwise, need to be considered, especially if these new forms of data are to be of direct value to policy makers and other data users.

Further, methods of record linkage so widely used in the government statistical community are based solely on numerical information and these will need to be extended to deal with mixed-media data. Consider for example, key variables for matching that involve related concepts but different media for representation in two different data bases. What does it mean to match records in such circumstances?

Finally, mixed media data will also pose new challenges for disclosure limitation methodology, as statistical agencies attempt to preserve the confidentiality of identifiable information provided by respondents. For example, unaltered images allow for the unique identification of respondents. Not releasing the images in some form however, takes away from the value of the multiple-media data sets for secondary analysis. Thus we need new approaches. One possibility might be to develop methods for simulating new images for release to replace the original ones, an image-based analogue to a general class of approaches I made for numerical data in the symposium here two years ago (see Fienberg (1996)). When surveys involve the collection of other kinds of information, such as soil and water samples issues and concerns with confidentiality are just as important.

To achieve a vision of restructured methods for survey data collection and analysis that would allow for exploiting the power of mixed media, extensive research efforts are required and these are expected to occur as part of an interdisciplinary effort involving computer scientists and statisticians, as well as the users of government statistical data bases.

5. SUMMARY

I began this presentation by recalling how far we had come in the past 20 years or so in terms of the use of model based methods for the analysis of surveys and censuses and the impact that technology has had on the tools available for survey work today. But such changes have occurred at a time when technology and society are changing at an even faster rate. Thus we should not be surprised that some of the finest survey and census work comes under regular challenge. I described a few of these challenges as well.

I have emphasized methods linked to technological advances, especially linked to computing technological developments. One of the big advances which most of us have not fully experienced is the joining of computing with cable television systems, something that will yield broad channel transmission of data and access to the type of multiple-media data sets I have described. I am not a believer in technology for technology's sake, but rather because I see computer technology as an enabler. Further, today's students of statistics think in terms of such technologies in very different ways than do most of us, and this leads to a different kind of methodological creativity than we saw even a decade ago.

My remarks this morning are part of a call to action, a call for statistical methodologists to begin to look ahead, not just to the solution of the problems described in the traditional methodological research which forms the core of the papers presented at this Symposium, but to the striking challenges posed by multiple-media data. As I have tried to explain, such data have already entered the domain of survey research, and I believe that they will take on increasing importance. The challenges that I have described are both technological and methodological and they offer exciting statistical work that should take us well into the new millennium. Statistical agencies ultimately will become engaged in the methodology, but they should begin exploring the technology soon, before the demand for such data puts them at a disadvantage.

ACKNOWLEDGEMENTS

The preparation of this paper was supported in part under a contract with Westat and the U.S. Bureau of the Census, and by a Research Grant No. REC-9720374 to the Center for Automated Learning and Discovery at Carnegie Mellon University. I thank Bob Boruch, Bill Eddy, Alicia Carriquiry, Tim Holt, and Sarah Nusser for reactions and suggestions as well as specific materials described here. Bill Eddy supplied the images included here as Figures.

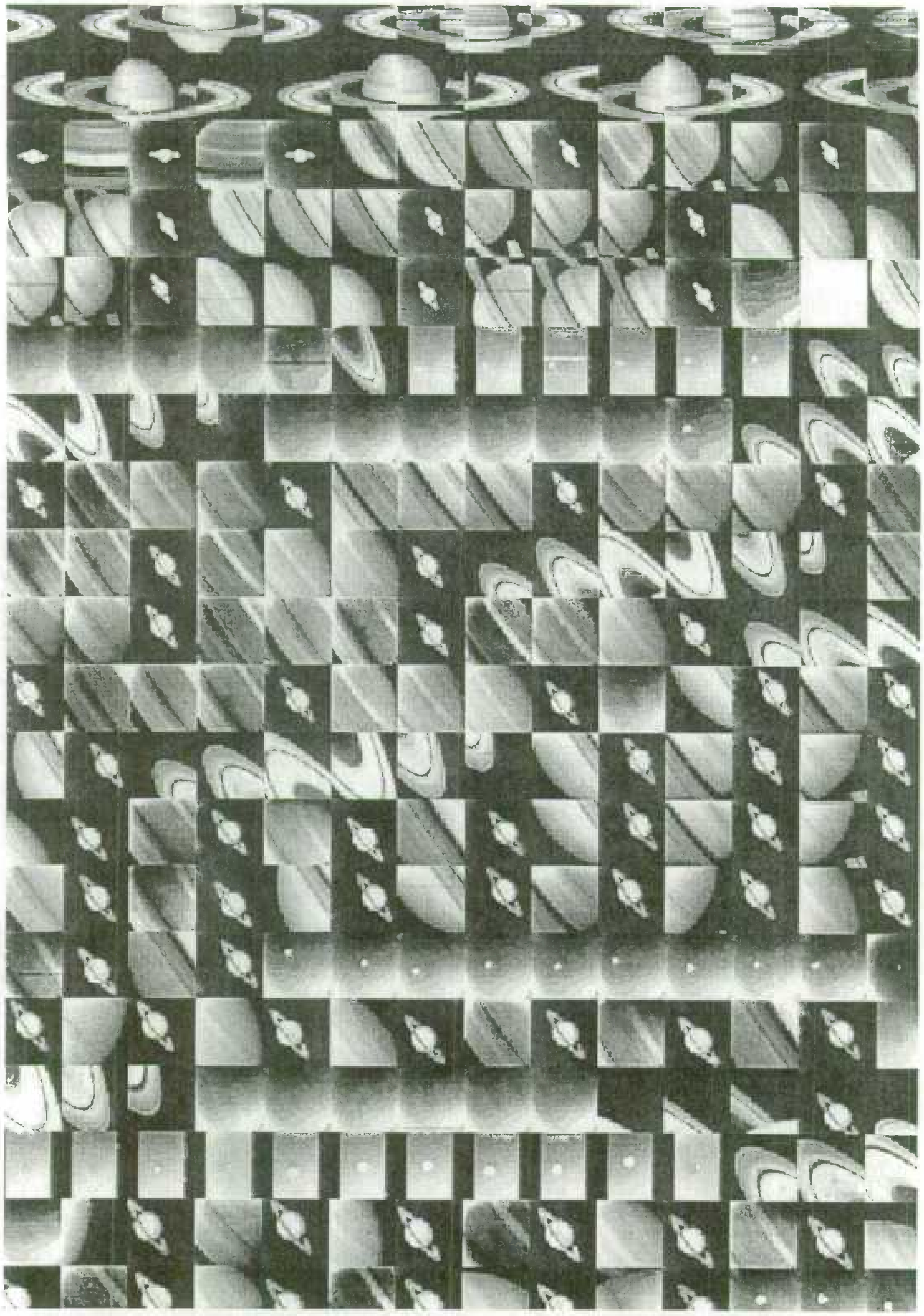


Figure 1. Extract of "contact sheet" of Voyager spacecraft images of Saturn used in Eddy-Mockus Interactive Icon Index system.

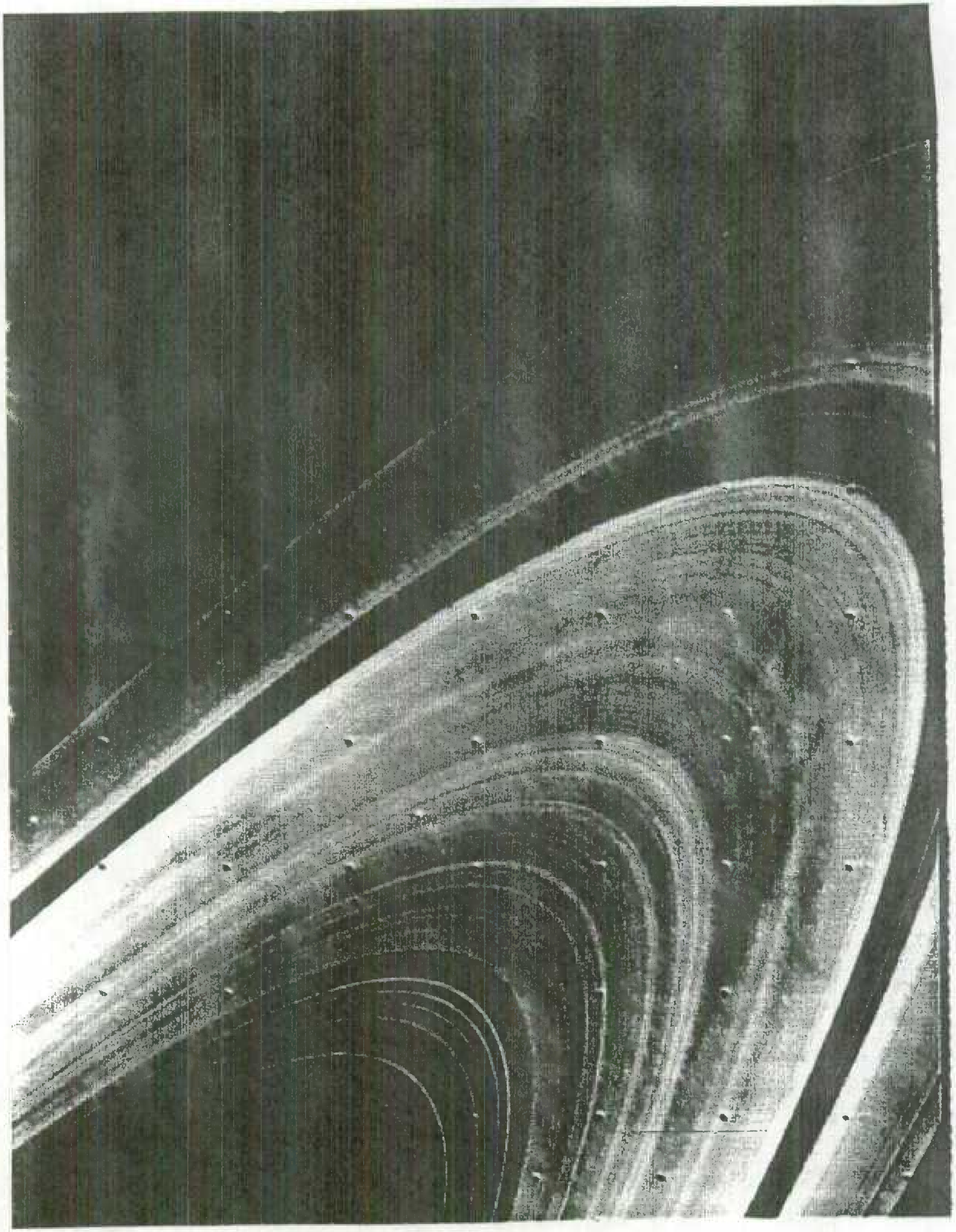


Figure 2. Voyager Image of Saturn linked to contact sheet displayed in Figure 1.

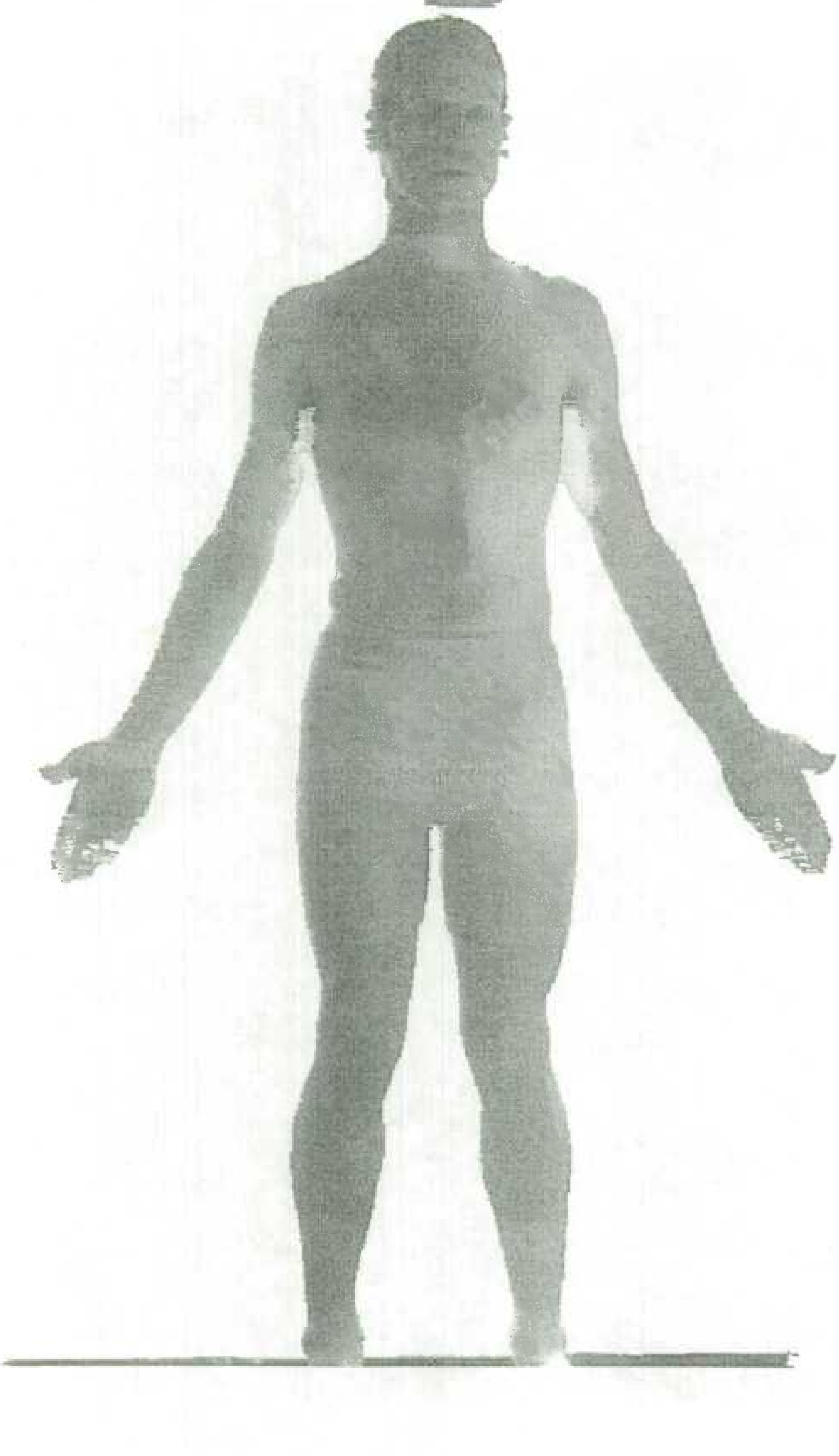


Figure 3. Two dimensional picture of young male from three-dimensional laser full-body scan generated in survey pretest by Anthropometric Research and Design Laboratories, Wright-Patterson Air Force Base.

REFERENCES

Advisory Committee on Science and Technology Statistics. (1997). Draft Framework for S&T Statistics. Working Group Report, Information System for Science and Technology Project, Statistics Canada, Ottawa.

Anderson, M., and Fienberg, S.E. (1997). Who counts: The politics of census taking. *Society* (Transaction), 34 (No. 3, March/April), 19-26.

Anderson, M., and Fienberg, S.E. (1998). *Who Counts? The Politics of Census Taking in Contemporary America: Adjustment and the 1990 Census*. New York: Russell Sage Foundation, in preparation.

Bollinger, C.R., and David, M.H. (1997). Modeling discrete choice with response error: Food stamp participation. *Journal of the American Statistical Association*, 92, 827-835.

Boruch, R.F., and Terhanian, G. (1996). 'So what?' The implications of new analytic methods for designing NCES surveys. In *From Data to Information: New Directions for the National Center for Education Statistics* (Eds. G. Hoachlander, J.E. Griffith, and J.H. Ralph). Washington, DC: National Center for Educational Statistics, 4-1-4-115.

Boskin, M.J., *et al.* (1996). *Toward a More Accurate Measure of the Cost of Living*. Final Report to the Senate Finance Committee from the Advisory Commission to Study The Consumer Price Index, Washington, DC.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. (Annex A to the Report by Jensen). *Bulletin of the International Statistical Institute*, 22, Supplement to Liv. 1, 1-62.

Brunsman, M.A., and Files, P.S. (1996). The CG dataset: Whole body surface scenes of 53 subjects (u). U.S. Air Force, Armstrong Laboratory, Report AL/CF-TR-1996-0160, Wright-Patterson AFB, Ottawa.

Bureau of Labor Statistics (1992). *BLS Handbook of Methods*. Volume II, Bulletin 2134-2 (December 1992) Washington, DC: U.S. Department of Labor.

Chesher, A. (1997). Diet revealed?: Semiparametric estimation of nutrient intake-age relationships (with discussion). *Journal of the Royal Statistical Society, Series A*, 160, 389-428.

Choldin, H.M. (1994). *Looking For the Last Percent: The Controversy Over Census Undercounts*. New Brunswick, NJ: Rutgers University Press.

Eddy, W.F., and Mockus, A. (1996). An Interactive Icon Index: Images of the Outer Planets. *Journal of Computational and Graphical Statistics*, 5, 100-111.

Edmonston, B., and Schultze. C., eds. (1994). *Modernizing the U. S. Census*, Washington, DC: National Academy Press.

Fay, R.E. (1984). Some properties of estimates of variance based on replication methods. *American Statistical Association Proceedings of Survey Research Methods Section*, 495-500.

Fay, R.E. (1992). When are inferences from multiple imputation valid?. *American Statistical Association Proceedings of Survey Research Methods Section*, 227-232.

Fienberg, S.E. (1978). Victimization and the national crime survey: Problems of design and analysis. In *Survey Sampling and Measurement* (Ed. K. Namboodiri). New York: Academic Press, 89-106.

Fienberg, S.E. (1994). An interview with Janet Norwood. *Statistical Science*, 9, 55-93.

Fienberg, S.E. (1994). Ethical and modeling considerations in correcting the results of the 1990 decennial census. In *Ethics in Modeling* (Ed. W.A. Wallace). New York: Pergamon Press, 103-144.

Fienberg, S.E. (1996). Taking uncertainty and error in censuses and surveys seriously. *Proceedings of Statistics Canada Symposium 95, From Data to Information, Methods and Systems*, 97-105.

Fienberg, S.E., and Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *Canadian Journal of Statistics*, 16, 135-151.

Fienberg, S.E., and Tanur, J.M. (1990). A historical perspective on the institutional bases for survey research in the United States. *Survey Methodology*, 16, 31-50.

Fienberg, S.E., and Tanur, J.M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review*, 64, 237-253.

Gould, S.J. (1997). *Questioning the Millennium: A Rationalist's Guide to a Precisely Arbitrary Countdown*, New York: Harmony Books.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 574-590.

Holmes, S.A. (1997). Tentative pact will allow census to test the sampling method. *New York Times*, November 1, 1997.

Jensen, A. (1926). Report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22, Liv. 1, 359-380.

Kiaer, A.N. (1897). *The Representative Method of Statistical Surveys*. Translation (1976), Central Oslo, Norway: Bureau of Statistics.

Malec, D., Sedransk, J., Moriarity, C.L., and LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Mulry, M.H., and Spencer, B.D. (1993). Accuracy of the 1990 census and undercount adjustments. *Journal of the American Statistical Association*, 88, 1080-1091.

Neyman, J. (1934). On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection (with discussion). *Journal of the Royal Statistical Society*, 97, 558-625.

Nusser, S.M., Carriquiry, A.L., Dodd, K.W., and Fuller, W.A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91, 1440-1449.

Robinette, K.M. (1995). Female anthropometry and the implications for protective equipment. *SAFE Journal*, 25, 35-45.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Scheuren, F. (1996). Administrative record opportunities in educational survey research. In *From Data to Information: New Directions for the National Center for Education Statistics* (Eds. G. Hoachlander, J.E. Griffith, and J.H. Ralph). Washington, DC: National Center for Educational Statistics, 9-1-9-29.

Smith, T.M.F. (1994). Sample surveys 1975-1990: An age of reconciliation? (with discussion). *International Statistical Review*, 62, 5-34.

Stasny, E.A. (1987). Some Markov-chain models for nonresponse in estimating gross labor flows. *Journal of Official Statistics*, 3, 359-373.

Stasny, E.A., Goel, P.K., and Rumsey, D.J. (1991). County estimates of wheat production. *Survey Methodology*, 17, 211-225.

Steffey, D.L., and Bradburn, N.M., eds. (1994). *Counting People in the Information Age*, Washington, DC: National Academy Press.

Stevenson, R.W. (1997). Republicans abandon plan for survey on I.R.S.. *New York Times*, November 11, 1997.

Stigler, J.W. (1996). Large-scale video surveys for the study of classroom processes. In *From Data to Information: New Directions for the National Center for Education Statistics* (Eds. G. Hoachlander, J.E. Griffith, and J.H. Ralph). Washington, DC: National Center for Educational Statistics, 7-1-7-27.

Tanur, J.M., ed. (1992). *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, New York: Russell Sage Foundation.

SESSION I-1

Innovative Sample Designs

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

PROJECT TO IMPROVE PROVINCIAL ECONOMIC STATISTICS

D. Royce[1], F. Hardy and G. Beelen

ABSTRACT

In October 1996, the Government of Canada and the governments of Nova Scotia, New Brunswick and Newfoundland signed an agreement to harmonize their federal and provincial taxes effective April 1, 1997. Statistics Canada will be the primary source of the data to be used in the formula for allocating the pooled revenues among the participating governments. In order to meet this new requirement, Statistics Canada needs to substantially improve the quality, reliability and detail of its provincial economic statistics. Therefore, over the next three years, Statistics Canada will be undertaking a major project , known as the Project to Improve Provincial Economic Statistics (PIPES). The PIPES implementation plan is based on a major transformation of the current model used for the collection and compilation of business surveys. The paper describes some of the major changes that will be implemented over the next three years. The Business Register will become the single frame for all business surveys, the enterprise will become the primary unit for data collection and analysis, and the sampling, estimation and survey processing for all annual surveys will be integrated in a coherent manner. A new annual census of approximately 8000 "complex" enterprises will be conducted to collect consolidated income statements and balance sheets to produce improved estimates by province and industry, and a major data quality measurement and quality assurance program will be implemented.

KEY WORDS: Economic statistics; Business surveys; Harmonization; Integration; Enterprise.

1. INTRODUCTION

The Project to Improve Provincial Economic Statistics, or PIPES for short, is one of the most important new undertakings of Statistics Canada of the past twenty-five years. In this paper we present a general description of the objectives and strategy of PIPES, and discuss many of the methodological challenges we face.

Section 2 begins with the background to PIPES – describing why it is needed and its goals and major elements. Section 3 describes the largest of these elements – the Unified Enterprise Statistics Program (UESP). We present the general principles of the UESP and give an overview of the UESP strategy. Section 4, which is the main part of the paper, focuses on the methodological challenges we are facing in the design of the UESP. In many cases, we still have more questions than answers. Finally, Section 5 concludes with a description of where we are in the project and our plans for the future.

2. BACKGROUND TO PIPES

PIPES is a direct result of an agreement between the Government of Canada and the provincial governments of Newfoundland, Nova Scotia and New Brunswick to harmonize the federal Goods and Services Tax with the provincial sales taxes into a so-called Harmonized Sales Tax (HST). The agreement was signed in the fall of 1996, and took effect on April 1, 1997. The HST is collected by Revenue Canada, and the pooled revenues are allocated among the participants using, in large part, the Provincial Economic Accounts of Statistics Canada. At the time the agreement was signed, however, all parties recognized that the quality of the provincial accounts needed improvement if they were to be used to support the HST allocation. As a result Statistics Canada has been funded to make the improvements needed to support this new and important use of the data.

The details of the allocation formula are exceedingly complex and run to several pages. Very briefly, however, the entitlement of a province equals that province's provincial tax rate (prior to harmonization) times a provincial tax base. This tax base has four components, namely a consumer base, a housing base, a business sector base, and a public sector base. Each base is essentially a sum of final expenditures for each commodity, excluding tax, times the taxable proportion of that commodity. While some of the data needed to calculate these bases come from Revenue Canada, over 90% of the total base is derived from Statistics Canada data.

The main objective of PIPES is to supply the data needed to support the formula. In simple terms, this means the measurement of the final sales of goods and services, on an annual, calendar year basis, by province and industry, by commodity and class of customer. For purposes of use in the allocation formula, these data are transformed into a set of Provincial Economic Accounts.

The Provincial Economic Accounts use data on both expenditures and incomes, and so improvements are planned in both areas. On the expenditure, or demand side, the major improvements include a larger and more frequent Survey of Household Spending, improvements to the Canadian Travel Survey, and a more regular Repairs and Renovations Survey.

However the vast majority of the changes to our program will come on the supply side. The major component is the

[1] Don Royce, Business Survey Methods Division, 11th Floor, R.H. Coats Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

development of a new integrated program that will eventually replace the existing program of annual business surveys. This new vehicle is known as the Unified Enterprise Statistics Program (UESP). Because the UESP will take a few years to become operational, however, a number of temporary enhancements are being made in order to achieve improvements starting right away.

This new integrated approach was chosen instead of a general scaling up of our existing program, for three main reasons:

- PIPES will inevitably result in an increase in response burden for Canadian businesses. We must be able to measure and to report on this burden, and this is much easier to do under an integrated approach.
- Data quality improvements are needed in four areas: (i) better consistency of the methods used across industries, (ii) better coherence of the data collected from different levels of the business, (iii) better coverage of industries, and (iv) better depth of information, in the sense of more content detail and larger samples. All of these objectives, but especially consistency, are easier to achieve with an integrated approach.
- The simultaneous redesign of our entire program of business surveys in a short period of three to four years would have spread our experienced resources too thinly.

3. THE UNIFIED ENTERPRISE STATISTICS PROGRAM

In developing the UESP, we have decided upon some basic integrating principles:

- All data for an enterprise (the highest level in the statistical hierarchy) and its establishments will be collected and analysed together, in order to ensure complete coverage of value-added.
- There will be a particular focus on the so-called complex enterprises (defined below), which although small in number account for over half of all economic production.
- We will aim for equal quality of data for all provinces.
- The Business Register (BR) will be the sole source of the frame for the UESP – in contrast to the current situation where not all surveys use the BR as the frame.
- We will use standard methodologies, both to make the resulting data more consistent and to make it easier to change in the future.
- In order to minimize response burden, administrative data will be used wherever possible instead of survey data.
- We will harmonize the concepts, definitions and questionnaires used in business surveys.
- We will have a formal program of Data Quality Measurement and Assurance for the UESP.

The UESP vehicle consists of four parts (see Figure 1), delineated by the type of enterprise (complex or simple) and the type of data (enterprise-level or establishment-level) required. Part 1 will consist of a census, or near census, of the approximately 8,600 complex enterprises. A complex enterprise is one which operates in more than one province, or more than one industry, or has more than one legal entity associated with it. Because of this, tax data do not provide the detailed breakdown by industry and province that is needed. Many, but not all, of the complex enterprises are large. The Part 1 questionnaire will collect primarily financial data at the enterprise level. Part 2 will be a census or near census of the roughly 60,000 establishments belonging to the complex enterprise. This part will collect financial and non-financial data at the establishment level. Part 3 will be a sample survey collecting primarily non-financial data at the establishment level for the simple enterprises. The last part is tax data. In the next few years Revenue Canada will be implementing the so-called "General Index of Financial Information", which will provide financial data, in electronic form, for all tax filers, both incorporated and unincorporated. Because simple enterprises operate in only one industry and province, tax data can supply virtually all of the financial data required for PIPES.

	Enterprise Data	Establishment Data
Complex Enterprises	Part 1	Part 2
Simple Enterprises	Tax Data	Part 3

Figure 1. Components of the Unified Enterprise Survey

4. METHODOLOGICAL CHALLENGES

The development of the UESP poses methodological challenges in many different areas. This section describes the major ones on which we are focussing as of late 1997.

4.1 Questionnaire Issues

First, our objective is to harmonize and simplify the number of questionnaires sent to businesses. Statistics Canada currently has about 100 annual business surveys involving over 700 different questionnaires, so this will not be a trivial task.

The UESP approach is to have one consolidated questionnaire at the enterprise level, plus industry-specific "schedules" at the establishment level. Standard concepts and question wording would be used across industries wherever possible, although we know that some variations will be required.

We also want to make it easier to complete our questionnaires by simplifying them, using concepts familiar to the respondent, and eventually by personalizing questionnaires. For example, in questions about commodities, we might only include those commodities reported in last year's report plus an "other" category in the questionnaire sent to a respondent.

4.2 Improvements to the Business Register

The quality of the frame is key to the success of the UESP. Efforts to reduce limitations of the Business

Register are currently underway on three fronts: coverage, frame variables and linkage to tax data.

First, the current Business Register excludes very small businesses. While these businesses account for a very small percentage of the economy overall, they can be important in some provinces and industries. Approaches need to be developed to supplement the coverage of the Register for these businesses. Second, recent improvements have been made to the Register by using more than one administrative source. While in principle these sources are linked through a common key known as the Business Number, in practice we are experiencing an increase in the amount of duplication on the Register as a result. Dead units, recent births, and out-of-date information may also be a problem, since the Register will be used in a number of industries for which there has been no survey feedback. Finally, the need to treat an enterprise and its establishments together makes the profiling of complex businesses a key activity for the UESP.

The quality of frame variables such as industry classifications and size measures is also important, since they are used for stratification. We are currently in the process of converting the Business Register from the 1980 Standard Industrial Classification to the new North American Industrial Classification System (NAICS), but the process will not be fully complete until the spring of 1999. In the meantime, we expect to have some growing pains due to the use of default NAICS codes. For size measures, the recent availability of new data sources linked through the Business Number offers great potential for the improvement of the size measures.

Samples chosen from the Business Register must be able to be linked to tax data, if the latter are to be exploited as a replacement for survey data. While these links are relatively good for incorporated businesses, links for unincorporated businesses need considerable improvement. Work is underway to improve these linkages using a variety of data sources.

4.3 Sampling, Collection and Response Burden Issues

Design of a sampling and collection strategy which provides good data while incorporating response burden concerns will be a major challenge. Response burden is a particular concern for small businesses, some of whom may not even be capable of providing the information we require. At present we are considering two or possibly three options.

One option would be to first stratify the population into three groups, based on the level of detail we felt they could provide. Large businesses would be subject to the full level of detail, medium-sized businesses would be subject only to an abbreviated questionnaire, and small businesses would not receive a questionnaire at all – the only data potentially available would be from tax. Separate samples, each of which might include sub-stratification by size, would be chosen within each of these strata. For the short form and tax data only strata, mass imputation would be used to impute for those variables on the long form but not collected on the short form or available from tax. This imputation would be based on models built from the larger businesses where the full data are available.

This approach has the strong advantage that it minimizes response burden where most important – for small businesses. However, the mass imputation of smaller businesses using models built from the larger businesses will inevitably introduce some degree of bias. The development of these mass imputation models may also be quite complex. The method would require good size measures and good correlation between the size measures and the variables of interest.

A second option is to use multi-phase sampling. For example, a large first-phase sample could be chosen and tax data obtained. A second-phase sample of these units could be asked to report the variables contained on the short form, and a third-phase sample could be asked to report the full detail. Different sampling rates could be used depending on the size class.

Information collected at earlier phases could be used as auxiliary information in estimation at later stages. This could reduce overall sample sizes and therefore response burden. The approach also has the advantage of being able to produce unbiased estimates. However it does imply a higher response burden for at least some small businesses that would receive a long form. Some form of sample rotation would probably be required to ensure that this burden was spread over the population. The sampling and weighting would also be more complex.

A third option is to use some combination of these techniques. For example, very small businesses might still be excluded from surveying, medium size businesses might use the full three phase approach, and large businesses might use a two-phase approach, consisting of tax data in the first phase and a long form in the second phase. In the future, once tax data become available in electronic form for all businesses, the tax phase sampling rate can become 100%.

Another important issue to be resolved is that of "equal data quality" for each province and territory. Equal coefficients of variation would result in high, and probably unacceptable, sampling fractions in the smaller provinces. This aspect will have to be factored into the allocation of the sample to provinces and territories.

The enterprise-centric approach will affect the design of the sampling and data collection. In the past, data at the enterprise and the establishment level have been collected by separate survey vehicles, with quite different sample designs. The UESP approach, which combines them into one vehicle, requires an approach that meets both needs. Many aspects of the sample design are affected, such as the choice of a sampling unit, the sample sizes at both the enterprise and establishment level, and the type of stratification and allocation used. Collection of data for both the enterprise and its establishments in a coherent fashion must also be addressed. Considerable investigation is required before the best approach can be identified. Should we attempt to collect establishment-level data from the enterprise? What kinds of consistency edits between the two levels should be applied? How flexible should we be in accepting roll-ups of establishment-level data to higher levels? These are a few of the questions that need to be answered.

Timing of sampling and data collection is also an issue. In Canada, fiscal years for incorporated businesses can end on any date. Experience suggests that it is worthwhile to collect the data a few months after the end of a business's fiscal year. However such a collection strategy affects the timing of sample selection. Should the sample be selected just once per year, to meet the earliest data collection date, even though the frame may not be as up to date as it could be later in the year? Or should we create a frame and sample from it several times a year, with the conceptual problems this creates?

We also expect that selective follow-up will be used in order to concentrate our collection efforts on the most important units, while at the same time reducing response burden on the smaller businesses. Some type of score function will be required for this purpose. How this score function should be linked to the sample design is an area that remains to be studied.

4.4 Processing Issues

There are also several issues in the area of processing. The challenge of harmonizing some 700 different questionnaires carries over to the harmonization of edit and imputation methods across industries. The need to treat enterprise and establishment-level data in a coherent fashion also raises issues such as how to use data from one level in editing and imputing data for another level. As mentioned earlier, the development of mass imputation methods to fill in the blanks for units intentionally not surveyed, or to allocate data from one level to another, will be required. Finally, data are required on a calendar year basis, although we intend to collect it from respondents or obtain it from tax on a fiscal year basis. Thus, methods will be required to transform the data to a calendar year basis.

4.5 Use of Tax Data

Increased use of tax data is fundamental to the UESP. At the present time, we only have access to a very limited number of variables for the entire universe. For certain programs and industries, we also select samples of tax data and capture more detailed information. Over the next few years, however, it is expected that we will have access to tax data from Revenue Canada in electronic form for 100% of the business population. Such tax data could serve a number of purposes:

- Tax data could form the basis of a supplementary frame for very small businesses, which are currently not well covered by the Business Register.
- Tax data could also serve as a source of frame information, such as measures of size.
- A major role for more detailed tax data would be to replace data collection for many enterprises, thereby reducing response burden.
- Tax data could serve as a supplementary source of information for imputation of non-response.
- Tax data could serve as auxiliary information for estimation and/or mass imputation.

4.6 Data Quality Issues

The improvement of the quality of our provincial economic data is at the heart of PIPES. The improvement in quality must be real, and it must be measurable. For this reason, we are instituting a formal data quality program to answer three main questions – what is the current level of data quality; what are the major areas requiring improvement, and how will we know if we have achieved better quality?

To answer the first question, the level of data quality in the current annual surveys program is being established by pulling together available information on data quality for reference years 1995 and 1996. This information will assist us both in identifying areas where quality improvements are needed, as well as identifying areas where data quality measures are lacking. Second, as described above, the major areas of quality requiring improvement are felt to be in the areas of consistency, coherence, breadth and depth. The integrated, enterprise-focused approach of the UESP is an attempt to address these weaknesses. Finally, the establishment of a formal program of data quality measurement, similar to programs in place for our other high-profile programs, will allow us to assess whether data quality is being improved. Over the next year we plan to develop a detailed strategy in this area.

5. CURRENT STATUS AND FUTURE WORK

The introduction of the UESP approach will require several years. For 1997 reference year, we are beginning with a pilot survey of seven industries not covered by current survey programs. This will allow us to try out the UESP approach while not disturbing existing programs. The pilot is not just a methodological test, however; the industries being covered represent real data gaps and it is planned to make use of the data collected by the pilot to improve the Provincial Economic Accounts. Mailout is planned for March 1998, with preliminary estimates available by December 1998.

In parallel, we are developing plans for transition of existing annual survey programs to the UESP approach for the 1998 and subsequent reference years. A number of factors must be considered in developing the transition strategy, such as the importance of the program to the PIPES objectives, the costs of making the transition, and the viability of current systems in light of the Year 2000 issue. A decision on the transition strategy is anticipated for late 1997.

In conclusion, PIPES represents a significant change to the way in which we conduct business surveys. The challenges ahead of us for the next several years are major ones. This is an exciting time in the business surveys area of Statistics Canada, and survey methodologists will have many opportunities to contribute to a fundamental and vital new program of Statistics Canada.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A SURVEY OF PROGRAM DYNAMICS FOR EVALUATING WELFARE REFORM

D.H. Weinberg[1], V.J. Huggins, R.A. Kominski and C.T. Nelson

ABSTRACT

The Personal Responsibility and Work Opportunity Reconciliation Act of 1996 eliminated the main United States welfare program, the Aid to Families with Dependent Children program, and replaced it with another program providing welfare support in the form of block grants to states. Part of that law directed the Census Bureau to field a survey, whose purpose is to collect the data necessary to evaluate the impact of this change. To carry out that directive, we are conducting a Survey of Program Dynamics (SPD). The SPD will simultaneously describe the full range of state welfare programs along with social, economic, demographic and family changes that will help or limit the effectiveness of the reforms. We will collect data for households previously interviewed from 1992-1994 or 1993-1995 by the Survey of Income and Program Participation for each of the six years from 1996 through 2001.

KEY WORDS: Welfare; Surveys; Evaluation.

1. INTRODUCTION

This paper will

- describe the need for a new survey focussed on providing the data necessary to adequately evaluate recent United States welfare reform legislation,
- describe the origin, purpose, status, and plans of the Survey of Program Dynamics, and
- discuss some technical issues we must resolve in the future.

2. WHY IS A NEW SURVEY NEEDED?

On August 22, 1996, President Clinton signed legislation passed by Congress, and the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 became Public Law 104-193. This comprehensive legislation has extensive implications for many programs. The law

- eliminates the open–ended federal entitlement program of Aid to Families with Dependent Children (AFDC),
- creates a new program called Temporary Assistance for Needy Families (TANF), which provides block grants for states to offer limited cash assistance,
- makes extensive changes to child care, the Food Stamp Program, Supplemental Security Income (SSI) for children, benefits for legal immigrants[2], and the Child Support Enforcement program,
- modifies children's nutrition programs,
- reduces the Social Services Block Grant, and
- retains child welfare and child protection programs.

The law also directs the U.S. Census Bureau to carry out a new survey to permit researchers to evaluate the impacts of the new law. Why would a new survey be needed?

Three kinds of information are critical to those investigating the effects of this welfare reform – process information (providing background descriptions), cross-section microdata (allowing comparisons of two points in time – "snapshots"), and longitudinal microdata (allowing pre-post analysis of the same individuals). All three kinds are needed to understand the full effects of the welfare reform legislation.

Process or descriptive information provides the context to interpret information about welfare recipients or former recipients. Examples of process information are the kinds of support services offered by welfare agencies (*e.g.*, child care, transportation vouchers, job search assistance), characteristics of the welfare agency itself (*e.g.*, cases per case worker), benefit levels, restrictions on client behavior (*e.g.*, whether a teenager must live with her parents), and so forth.

There will be several sources of such information. First is the information that states must report about their programs to the U.S. Department of Health and Human Services (DHHS). These reports are likely to have only the minimum necessary to satisfy the requirements of the legislation and therefore will probably be insufficient on their own for research purposes. Nevertheless, one can use this basic "tracking" information to tell some basic stories. Two other sources of descriptive information seem more promising, however. The first is an effort, funded by the U.S. Census Bureau through DHHS, taking place at the University of Wisconsin. The project team will attempt to describe all state programs, both pre-and post-reform,

[1] Daniel H. Weinberg, HHES Division, U.S. Bureau of the Census, Washington, D.C. 20233-8500.

This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect the views of the Census Bureau or the U.S. government.

[2] Since changed by the Balanced Budget Act of 1997.

along a set of common dimensions (descriptive factors). These would then become explanatory variables in investigations of outcomes using survey data. Also part of this project is a pilot study to see if such data can be collected at the county level. A parallel effort is underway at the Urban Institute (UI) as part of their Assessing New Federalism Project (ANFP).

Cross-section microdata can be and has been used to evaluate the effects of program changes. Most typically, researchers compare average characteristics of a population group (*e.g.*, the percent of welfare recipients working) at two points in time. Two sources of cross-section microdata will be available – the March Current Population Surveys (CPS), and the National Survey of American Families (NSAF), being conducted by Westat as part of the ANFP project in 1997 and possibly again in 1999 or 2000[3]. Because of the many variants of welfare and new forms of household support established by the states as they revise their assistance programs, survey organizations will have to make changes to existing survey questions to collect the relevant data.

The analysis of longitudinal microdata is the preferred approach when social experiments cannot be used to evaluate program changes, as with a nationwide change like the 1996 welfare reforms. Analysts use pre-reform characteristics of a population to control for preexisting differences among households when evaluating post-reform outcomes for the same people. This may be as simple as examining changes in employment for specific demographic groups, or as sophisticated as multiple regression that takes account of self-selection and sample attrition. Only two sources of longitudinal data will have large enough samples to analyze – the 1996 panel of the Survey of Income and Program Participation (SIPP), and the new Survey of Program Dynamics (SPD) – the new survey directed by the welfare reform law.

While the SIPP has a large sample size and will follow households for up to four years, it suffers from the deficiency that data collection began in April-July 1996, while the reform took effect on October 1, 1996. One could argue that four months of pre-reform information is sufficient for many analyses, and the SIPP does try to collect retrospective program participation information. Some, however, are skeptical, particularly those analysts who need a longer pre-reform period to accurately measure some initial condition, and particularly because many (if not most) states had already begun to make changes under federal program waivers well before the beginning of the SIPP panel[4].

On the other hand, the SPD, based on a sample of households first interviewed in February-May 1992 or 1993 to be followed until 2001, will provide a convincing set of baseline data, assuming as we must that differential attrition will not vitiate the usefulness of the data collected. The rest of this paper describes the SPD in more detail.

No one source of information will be complete. A full picture of the effects of welfare reform will emerge only after many years and complementary studies using these different sources.

3. ORIGIN AND PURPOSE OF THE SURVEY OF PROGRAM DYNAMICS

Why would a new longitudinal survey that focuses on welfare issues be needed? For particular agencies, a series of focussed single-purpose surveys or experiments can serve many of their specific program evaluation needs. But if the research community were to rely solely on highly focussed data collection, there would inevitably be major gaps. Only an omnibus data collection vehicle can provide the basis for an overall evaluation of how well welfare reforms are achieving the aims of the Administration and the Congress. This requires a survey that casts a wide net, one that simultaneously measures important features of (1) both reformed and unchanged welfare programs, and (2) other important social, economic, demographic and family changes that will either help or limit the effectiveness of the reforms. Further, ideally such a survey should be in place before reforms are effective to allow adequate assessment of baseline circumstances.

Several years before the passage of the actual legislation, DHHS and the U.S. Department of Agriculture (USDA. responsible for the food stamps program) invested substantial resources in having the Census Bureau develop such a survey. They hoped they could fund and field such a survey to meet their needs to understand the effects of anticipated public policy changes on the population. In these planning activities, several design features emerged as essential. The survey should:

- Measure
  - program eligibility and participation for the full range of welfare programs;
  - money income, in-kind benefits, and services received from programs;
  - employment, earned income, and income from other economic sources;
  - family composition; and
  - child outcomes including key features of the environments of children (because reforms may have positive or negative consequences for children through these intervening mechanisms);
- Be a large, longitudinal, nationally representative study that measures changes in each of these areas and allows the identification of interrelationships linking these changes;
- Include baseline data for a period before the initiation of reforms;
- Continue to collect data throughout the period of reform to monitor the process of change; and
- Collect data for the period after the states implement the reforms.

[3] The NSAF has quite respectable samples in 13 states, with a supplementary sample in the balance of the U.S. The survey has a sample size of about 35,000 households, with low-income households oversampled; the interview mode is computer-assisted telephone interviewing. A major focus of the survey is health outcomes. For more details on the NSAF, see Brick *et al.* (1997).

[4] More details about the SIPP can be found in U.S. Bureau of the Census (1991); a third edition is currently being prepared.

Section 414 of the welfare reform law specifically directs (and funds) the Bureau of the Census to:

> continue to collect data on the 1992 and 1993 panels of the Survey of Income and Program Participation [SIPP] as necessary to obtain such information as will enable interested persons to evaluate the impact of the amendments made by Title I of the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 on a random national sample of recipients of assistance under State programs funded under this part and (as appropriate) other low income families, and in doing so, shall pay particular attention to the issues of out-of-wedlock birth, welfare dependency, the beginning and end of welfare spells, and shall obtain information about the status of children participating in such panels.

To comply with this directive, the Census Bureau is carrying out the Survey of Program Dynamics (SPD) with two primary goals:

– Provide information on spells of actual and potential program participation over a ten-year period, 1992 to 2001, and

– Examine the causes of program participation and its long-term effects on the well-being of recipients, their families, and their children.

The data already collected in the 1992 and 1993 SIPP panels will provide extensive baseline (background) information from which to figure out the effects of welfare reform. SIPP is a longitudinal survey of households, interviewed at least nine times at four-month intervals, and followed if they moved. The SIPP collects more detailed data than any other national survey on program eligibility, access and participation, transfer income, and in-kind benefits. Coupled with an extensive array of economic and demographic data (*e.g.*, employment and job transitions, income, and family composition), the SIPP will serve to characterize the pre-reform situation of households quite well.

Further, the Census Bureau worked closely with policy agencies to develop and field topical modules that enhance the value of the basic SIPP data. Modules of special interest here include those on (1) education and training, (2) marital, fertility, migration, and program participation histories, (3) family relationships within the home, (4) work schedules, child care, child support, support for non-household members, (5) medical expenses and use of health care services, and (6) child well-being.

By interviewing the same households in the SPD, analysts would then have data for the baseline pre-reform period, the reform implementation period, and the medium-term post-reform period. Researchers require these data to assess short-term and medium-term consequences and outcomes for families and individuals. The use of both panels will also double the size of certain groups of interest, subject of course to our ability to recontact households in the two panels and their willingness to participate. (Because the funding provided is not sufficient to interview all households in both panels past 1997, we will subsample after 1997; see section IV.)

The topics that the SPD will cover are an extension of those covered by the SIPP, but placed in an annual context using guidance from such annual surveys as the March supplement to the Current Population Survey, the Panel Study of Income Dynamics, and the National Longitudinal Surveys.

4. SPD STATUS AND PLANS

Current plans are for data to be collected for each of the six years from 1996 through 2001. This will provide panel data for ten years (1992-2001) when combined with the 1992 SIPP panel data (and nine when combined with the 1993 SIPP panel data). Our original plans were to have an instrument ready to field concurrently with welfare reform. Because President Clinton vetoed the legislation twice during 1995, we put our plans on hold. Consequently, we were unable to pretest the SPD questionnaire and could not field the survey we had designed in 1997.

Nevertheless, we felt it critical to fill the data gap between the end of the SIPP observations and the start of the basic SPD observations. To do so, we designed the SPD with three fundamental sections:

(1) the "bridge" survey which will provide the link between the 1992 and 1993 panels of the SIPP and the SPD;

(2) the 1998 SPD which will use the core instrument already developed to collect annual retrospective data starting in 1998; and

(3) the 1999 (and later) SPD which will include a child well-being module starting in 1999; its content may vary from year to year.

SPD "Bridge" Survey

It was critical to collect income and program participation data in spring 1997 for calendar year 1996 from as many of the 1992 and 1993 SIPP households as we could find, as too long a gap ran the risk of losing too many households and missing too much going on in their lives. Data for 1996 were collected in April-June 1997 by administering a modified version of the annual March 1997 Current Population Survey (CPS) demographic supplement, with a few new questions designed to collect summary 1995 data for the 1992 SIPP panel (who were last interviewed in January 1995).

Eligible for the SPD sample were all SIPP persons interviewed in the first wave of the 1992 and 1993 SIPP panels and still being interviewed at the end of the panel. We decided not to try to find all persons in the 1992 and 1993 panels who left the survey (attrited) before the end of the SIPP because of the difficulty and cost involved in trying to find those people (who had already declined to participate even after repeated attempts to interview them) and because we felt that most analysts would need as much baseline data as possible. Use of population controls will reweight the remaining sample cases to represent the U.S. population.

Finding people who move is critical to the success of any longitudinal survey, particularly one as focussed on the low-income population as SPD. Naturally, SIPP has developed many time-tested procedures that will help. But the automated questionnaire instrument must permit such tracking as well. Luckily, the CPS implemented a "mover module" in January 1997 to track people leaving formerly interviewed CPS households. This was crucial to the SPD especially given the time that had elapsed since the last interview, though the CPS approach did not fully meet the needs of the SPD. We tracked most, though not all, movers and we will attempt to interview in 1998 those that we know about but could not interview in 1997. We also tested the use of a $20 monetary incentive for low-income households in an attempt to reduce nonresponse to the Bridge survey; as the Census Bureau has shown that such an incentive was successful in reducing nonresponse to wave 1 of the 1996 SIPP panel[5].

1998 SPD

During the first half of 1997, the University of California at Berkeley translated the instrument developed in 1995[6] into computer code; UC-Berkeley is the developer of the CASES authoring language used for computer-assisted Census Bureau surveys. We carried out a pretest in October 1997 using 400 retired 1996 CPS households in four of our regional office locations. From this test we will have a good idea of how well the instrument does in an operational environment. We also will test the use of a self-administered adolescent questionnaire using audio cassettes to obtain information from youths 12 to 17 years old. Preliminary indications are that we may have to shorten the questionnaire to fit within our time constraint.

Using the fully developed computer-assisted personal interview instrument, annual data collection will occur once each year in May and June, with annual recall for the preceding calendar year. The survey will include a set of retrospective questions covering 1997 for all persons aged 15 and older in the household. The topics covered are:

- Basic demographic characteristics, including
  - educational enrollment and work training,
  - functional limitations and disability, and
  - health care use and health insurance;
- Basic economic characteristics, including
  - employment and earnings,
  - income sources and amounts,
  - assets, liabilities, and program eligibility information, and
  - food security;
- Information about children, including their
  - school enrollment and enrichment activities,
  - disability and health care use,
  - contact with absent parent,
  - care arrangements, and payment of child support of their behalf; and
- two self-administered questionnaires,
  - a short question sequence for adults focussing on marital relationship and conflict and a depression scale, and
  - a relatively lengthy questionnaire for adolescents aged 12 to 17 focussing on such issues as family conflict, vocational goals, educational aspirations, crime-related violence, substance use, and sexual activity, (developed in collaboration with the Child and Family Research Network)[7].

1999 SPD and later

Work has begun on identifying the topics for a child well-being module to be asked in 1999 or later. We plan to focus on elements that allow analysts to measure changes from pre-reform periods or that illuminate other mechanisms affecting outcomes. One possibility is to collect data on where the children have lived and the reasons for any absence from the parents. Also under investigation are question variants to address the changing nature of welfare programs in the 50 states and the District of Columbia.

5. TECHNICAL ISSUES

This section deals with four technical issues that affect the administration and usefulness of the survey – the need for subsampling due to budget constraints, weighting and database development, the collection of supplementary data, and the use of administrative records.

Subsampling. It is clear we cannot interview all households in the 1992 and 1993 SIPP panels in 1998. The response rate to the 1997 Bridge survey was good, given the time that had elapsed since the prior interview – 81.7 percent – yielding a sample of 30,125 interviewed households. The budget for the survey, $10 million per year, will allow us to complete interviews with about 17,500 households in 1998, given the projected length of the interview. Our current plans are to

- Sample with certainty
  - all households with income less than 150 percent of poverty, and
  - all households with incomes between 150 percent and 200 percent of poverty with children;
- Subsample at 80 percent
  - all households with incomes above 200 percent of poverty with children;
- Subsample at 50 percent
  - all households with incomes between 150 percent and 200 percent of poverty without children; and
- Subsample at 26 percent
  - all households with incomes above 200 percent of poverty without children.

Subsampling will be based upon household characteristics as of the Bridge survey.

[5] See James (1998). Evaluation of the effectiveness of the SPD incentive is underway.

[6] For more information on the design process for SPD, see Hess and Rothgeb (1997).

[7] We have not yet finalized this questionnaire.

Weighting and database development. Three objectives will influence our thinking on SPD data products and weighting. We want to:

- Provide longitudinal data to evaluate the effects of welfare reform;
- Release a data product as soon as possible after collecting the 1997 Bridge data; and.
- Focus our scarce resources on just the products that our users need most.

The first product the Census Bureau will release is a public use microdata file. This will include data from the 1997 Bridge data, longitudinal weights to post-stratify up to a January 1993 cohort, and codes to link the SIPP and SPD Bridge data. We will release it as a research file with appropriate caveats. The weights will be crude, but should suffice for preliminary research purposes. Researchers always have the option of making additional weighting or imputation adjustments as they deem necessary for their specific analyses.

We realize that users of the longitudinal data might have a hard time figuring out how to use data from three separate surveys (SIPP, CPS, and SPD) simultaneously in a longitudinal analysis. Our challenge is therefore to create a longitudinal data set with annual data from all three survey instruments (SIPP, CPS, and SPD) in a consistent format[8]. Our current plans are as follows. The Census Bureau will release two files each year after 1998 and later SPD data have been collected, processed, and weighted. The files will include and provide appropriate weights for households responding to the latest interview. The first file will include (1) the SIPP 1992 and 1993 panel data covering 1992-1994 and 1993-1995 respectively, (2) the 1997 bridge data covering 1995-1996 (1992 panel) or 1996 only (1993 panel), and (3) the 1998 and later SPD data as originally collected and edited (covering 1997 and later). The second file will attempt to create a consistent set of annual measures for each year of data (a "common format" file), to simplify the analyst's chores; this file will probably use the "least common denominator" – the CPS – as the common data format[9].

These longitudinal products are conceptually the same type of products that we have issued for the SIPP since its inception. Usually, one defines a longitudinal cohort as all the people interviewed (or for whom data was imputed) in every interview within the period of interest. For example, the 1993-1997 SPD longitudinal cohort will be all those people interviewed from February 1993 through the 1998 SPD interview (May-June 1998). The control date will be as of the beginning of the cohort period (February 1993 for this example). This creates a nationally representative longitudinal cohort for the population as of the beginning of the cohort period. We expect the weighting adjustments will compensate for differential nonresponse.

Supplementary data. As noted earlier, the University of Wisconsin will create a complementary data base of state and county welfare program characteristics to match to the SPD data. County-level matches must remain confidential, and researchers would have to work on that matched data set at the Census Bureau under Special Sworn Employee status to maintain respondents' confidentiality.

Administrative Records. One final hope is that administrative records can enhance the SIPP and SPD survey data to provide an even broader and long-run picture of the pre- and post-reform economic situation of the survey households. For example, Summary Earnings Records from the Social Security Administration for all respondents providing Social Security Numbers (SSN's) could provide a good way of validating and extending the survey earnings reports (as would matches to income reports provided to the Census Bureau from Internal Revenue Service tax records).

One potentially useful development would be to establish a nationwide tracking system for welfare recipients based on SSN; such a system could be used to facilitate enforcement of the five-year lifetime assistance limit in the new legislation. We could then append location information and possibly benefits received to the survey data for welfare participants, even for those attriting from the sample.

A few other alternatives present themselves. If the SPD can collect employer name and address successfully, a match to the Standard Statistical Establishment List can provide the key to links with business data for those in the sample who are working. If states collect and keep consistent caseload data using SSN's, and their laws permit sharing such information with the Census Bureau, that's another possibility for enhancing the data files. Finally, we could match summary financial data about welfare and related expenditures of local and state governments using the household's residential location. (Of course, researchers must use confidential information at Census Bureau headquarters or one of its Research Data Centers.)

6. CONCLUDING REMARKS

The opportunity to do the Survey of Program Dynamics is an exciting one for the Census Bureau. It is also a sensible investment for the government as it builds on the investment it already has in the Survey of Income and Program Participation. Yet the challenges are daunting, especially given the length of time we will need to follow households, and as household cooperation with government surveys continues to decline.

ACKNOWLEDGEMENTS

The authors would like to acknowledge and thank Evan Davey and Stephanie Shipp for their comments and suggestions and the SPD Team for doing all the work described herein; they bear no responsibility for any errors than remain.

[8] The microdata from SIPP surveys are already available on-line at the Census Bureau's web site (www.census.gov) through the "Surveys on-Call" program; plans are underway to provide the data as well through FERRET (the Census Bureau's Federal Electronic Research and Review Extraction Tool, which already provides the CPS microdata to users).

[9] Researchers will be able to analyze drop outs by comparing the prior year's data file with the current one.

REFERENCES

Brick, J.M., Waskberg, J., Shapiro, G., Flores-Cervantes, I., Bell, B., and Ferraro, D. (1997). Nonresponse and coverage adjustment for a dual frame survey. To appear in *Proceedings of Statistics Canada, Symposium 97: New Directions in Surveys and Censuses*, Ottawa, 1997.

Hess, J., and Rothgeb, J. (1997). Measuring the impact of welfare reform: issues in designing the survey of program dynamics questionnaire. To appear in *Proceedings of Statistics Canada, Symposium 97: New Directions in Surveys and Censuses*, Ottawa, 1997.

James, T.M. (1996). Results of the wave 1 incentive experiment in the 1996 survey of income and program participation. *1997 Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1998.

U.S. Census Bureau, *Survey of Income and Program Participation Users' Guide,* Second Edition, Washington DC, 1991.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A SOLUTION TO THE DESIGN AND IMPLEMENTATION OF A FAST-TRACK SURVEY: TWO-PHASE SAMPLING

M. Brodeur[1], W. Jocelyn and J. Trépanier

ABSTRACT

A new Retail Commodity Survey was launched by Statistics Canada in January 1997, to obtain more information on commodity by sales in Canada. The survey is based on a two-phase sample design, where the first phase is the Monthly Retail Trade Survey, in place since 1988. Information from the first phase is used in all steps of the commodity survey in order to maximize the efficiency of the design: Multivariate sample allocation, sample selection, maintenance and imputation. Some development has been done to provide explicit variance estimation methods for a two-phase stratified sampling design where the second phase sample is selected from the restratified first phase sample.

KEY WORDS: Sampling; Two-phase; Restratification; Multivariate; Estimation.

1. INTRODUCTION

Today, tight budgets and short deadlines are forcing statisticians to devise innovative sample designs. Survey methodologies must be tailored to achieve a certain degree of efficiency within the limits imposed by such constraints. The Retail Commodity Survey (RCS) presented just such a challenge. It was designed and implemented over a period of one year. Its purpose was to collect detailed information about retail commodity sales in Canada. Since total retail sales were already being measured by another survey, the Monthly Retail Trade Survey (MRTS), we decided to use a two-phase sample design for the RCS in order to take full advantage of the MRTS's information and infrastructure. By taking this approach, we were able to develop the RCS more quickly, more economically and, from a statistical perspective, more efficiently.

In this paper, we present a methodological solution to the problem of designing and developing a fast-track survey: two-phase sampling. A brief description of fast-track surveys is provided in section 2, while section 3 contains an overview of the MRTS. Section 4 provides a detailed picture of the RCS's methodology, including the sampling plan, the sample allocation method, the edit, imputation and estimation strategies.

2. FAST-TRACK SURVEYS

The RCS's main purpose is to measure the distribution of retail sales by commodity group. A quarterly survey, the RCS satisfies requirements stated by data users both within and outside Statistics Canada. Similar surveys were conducted in 1974 and 1989, but in addition to being sporadic, they did not produce all the desired results. Nevertheless, the results of the 1989 survey helped in the development of the RCS's sampling plan.

A two-phase sample design is advantageous in many respects. First, it makes it possible to fast-track the introduction of a new survey. In this case, the time limit was one year. Second, the first phase information can be used to maximize efficiency in a number of areas, including the selection of the same respondents, the use of auxiliary information in sample allocation, edit, imputation and estimation, and the use of existing systems and staff.

Two-phase sampling was introduced by Neyman (1938) and was later treated at some length by Cochran (1963) under the name "double sampling". Since then, it has been studied by other researchers, including Särndal, Swensson and Wretman (1992) and Hidiroglou and Särndal (1995). However, none of them deals specifically with the case of a stratified design in which the first phase sample is completely restratified to make the second phase sample design as efficient as possible. That is what makes the RCS design innovative.

3. OVERVIEW OF THE MRTS: FIRST PHASE OF THE RCS

The Monthly Retail Trade Survey essentially measures retail sales by trade group (three-or four-digit groups based on the 1980 Standard Industrial Classification (SIC)) for each province and selected census metropolitan areas (CMAs). It was last redesigned in 1988. The sample is selected from Statistics Canada's Business Register (BR). The target population consists of statistical companies with statistical locations identified in the BR as retailers. Some 16,000 companies are interviewed each month. The population is stratified by province, territory, selected CMAs and trade group. Each combination of trade group and geographic area forms a stratum. Each stratum is divided into three substrata by size: one take-all stratum and two take-some strata, one composed of medium-sized firms and

[1] Marie Brodeur, Chief, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

the other of small companies. The take-all strata include all companies with a complex structure, *i.e.*, companies that operate in more than one trade group or geographic area or have a gross business income (GBI) above a certain limit. Other companies are assigned to one of the two take-some strata on the basis of their GBI. The target coefficient of variation of sales is 1.5% at the national level, 2.5% at the provincial level and 3.5% at the trade group level. Sample allocation is by the square root of GBI.

The sample is partially rotated each month to lighten the response burden and keep the response rate high. The population within each take-some stratum is randomly divided into equal-size clusters or panels. The number of panels is determined by the sampling fraction computed at the time of allocation and by the number of months a unit remains in the sample and outside the sample. A subset of the panels is selected for the sample. Rotation involves systematically removing one panel from the sample each month and replacing it with a new panel. The sample is rotated only in the take-some strata. Each month, births are systematically added to the panels. The expected number of births in the sample is thus reached.

4. RETAIL COMMODITY SURVEY

4.1 Second Phase Sampling Plan

In the second phase of sampling, information from the first phase is used to restratify and allocate the second phase sample. It is important to note that the second phase sample is a subset of the first phase sample. A unit that belongs to the second phase sample must, therefore, be part of the first phase sample. This section deals primarily with stratification, the sample allocation method, and rotation of the RCS sample.

4.1.1 Stratification of the RCS

The frame from which the RCS sample is drawn is the set of companies in the MRTS sample. As in the first phase, the sampling unit in the second phase is the statistical company. Using the latest information from the MRTS, the first phase sample is restratified by trade group, province and company size. For the purposes of stratification, each company is assigned a dominant province and a trade group on the basis of sales volume.

MRTS sales were used in determining the company size substrata. For operational reasons, however, that variable was not available for the entire sample. Therefore, sales had to be modelled using GBI, which was available for all units in the population. For consistency, the predicted sales provided by the model below were used to stratify the first phase sample. The following simple regression model was used:

$$V_i = \beta_0 + \beta_1 \,\mathrm{GBI}_i + \varepsilon_i$$

where V_i and GBI_i represent the sales and gross business income of company *i*. β_0 and β_1 are the model's usual parameters. The model's parameter estimates were used to predict sales for companies whose sales were unavailable in the first phase sample.

4.1.2 Sample Allocation

The RCS is intended to provide sales estimates for many commodity groups. Consequently, the sample allocation has to be multivariate. Since there is no conventional solution to the problem of optimal multivariate allocation when the survey is using a stratified two-phase sampling plan, an existing method had to be adapted to suit our sampling plan. The method chosen in this case is a modification of the Bethel (1992) algorithm. It is summarized below. For more details, see Jocelyn and Brodeur (1996).

Let ***K*** be the commodity groups. We want to estimate $\boldsymbol{T_k}$, the sales for a commodity group ***k***. Consider the following double expansion estimator for a two-phase sampling plan:

$$\hat{T}_k = \sum_{i=1}^{N} \sum_{h=1}^{H} \sum_{g=1}^{G} \frac{N_h}{n_h} \frac{M_g}{m_g} z_i z_i^{(2)} a_{ih} a_{ig}^{(2)} y_{ik}.$$

The variance of $\hat{T}_k$ is given by:

$$V[\hat{T}_k] = V_1 E[\hat{T}_k] + E_1 V[\hat{T}_k] = V_1 + V_2$$

with

$$V_1 = V_1 E[\hat{T}_k] = \sum_{h=1}^{H} \frac{N_h}{n_h} (N_h - n_h) S^2{}_{hk}$$

$$S^2{}_{hk} = \left[\sum_{i=1}^{N} \left[a_{ih} y_{ik} \right]^2 - \frac{\left[\sum_{i=1}^{N} a_{ih} y_{ik} \right]^2}{N_h} \right],$$

$$V_2 = E_1 V[T_k] =$$

$$E\left[\sum_{g=1}^{G} \left(\frac{M_g (M_g - m_g)}{m_g (M_g - 1)} \right) \left\{ \sum_{i=1}^{N} \left(\sum_{h=1}^{H} \frac{N_h}{n_h} z_i a_{ih} a_{ig}^{(2)} y_{ik} \right)^2 - \frac{\left(\sum_{i=1}^{N} \sum_{h=1}^{H} \frac{N_h}{n_h} z_i a_{ih} a_{ig}^{(2)} y_{ik} \right)^2}{M_g} \right\} \right].$$

where y_i is the first phase value of the variable of interest for unit *i*. H is a first phase stratum. a_{ih} is an indicator variable whose value is 1 if unit *i* is in stratum *h*, and 0 otherwise. Hence we can write $N_h = \sum_{i=1}^{N} a_{ih}$. z_i is an indicator variable whose value is 1 if unit *i* is in the first phase sample, and 0 otherwise. The sample size for the *h*-th stratum is therefore $n_h = \sum_{i=1}^{N} z_i a_{ih}$. N is the size of the population. G is a second phase stratum. $a_{ig}^{(2)}$ is an indicator variable whose value is 1 if unit *i* is in second phase stratum

g, and 0 otherwise, and $z_i^{(2)}$ takes the value 1 if unit i is in the second phase sample, and 0 otherwise. We can therefore write $M_g = \sum_{i=1}^{N} z_i a_{ig}^{(2)}$ and $m_g = \sum_{i=1}^{N} z_i^{(2)} a_{ig}^{(2)}$.

Hence the sample allocation problem can be solved by minimizing:

$$C = E_1 \left[\sum_{h=1}^{H} c_h n_h + \sum_{h=1}^{H} \sum_{g=1}^{G} c_g m_g \right]$$

provided that $\mathrm{CV}(\hat{T}_k) \leq \mu_k$ $k=1,2,3...,K$, where c_h and c_g are the unit costs of first phase and two respectively, and μ_k is the desired coefficient of variation. The following is the procedure for adapting Bethel's algorithm to the problem of second phase allocation where both first phase and second phase sample sizes must be determined at the same time. We first apply the algorithm to the H first phase strata, using V_1 as the variance. We compute the value of V_2 by replacing the corresponding m_g. We then apply the algorithm to V with V_2 modified. Finally, we calculate the coefficient of variation (CV) using the optimal values. If we fail to achieve the desired CV, we start over. We can show that this procedure also preserves the convergence properties of Bethel's algorithm. In our case, since the first phase sample sizes were predetermined, we simply applied the algorithm to modified V_2.

The CVs from the 1989 survey were used in applying Bethel's algorithm. The final sample size produced by the algorithm was about 10,000 units for a CV of sales of 7% at the Canada level for each major commodity group.

4.1.3 Sample Selection and Rotation

As mentioned earlier, the MRTS sample is made up of a subset of panels. The sample is partially rotated every month by replacing one of the panels. For the initial selection of the RCS sample, we ignored the panel structure of the first phase sample; this approach streamlined the process considerably. Since each first phase panel can be regarded as a simple random sample of the MRTS sample, the set of all first phase panels is also a simple random sample. Consequently, if we wish to draw a simple random subsample from that sample, we can do so without regard for the panel structure. That is what we did for the RCS.

To take the MRTS rotation into account, every month we select a subsample of units from the new MRTS panel. Since we assume in our estimation process that simple random sampling is used, we examined the effect that deviating slightly from that assumption would have on the estimates. The results showed that the effect was virtually non-existent.

4.2. Data Collection

As previously mentioned, the RCS sample is a subsample of the MRTS sample. MRTS data are mostly collected by telephone. The RCS's unit of collection is the same as the MRTS's. For these reasons, it was advantageous to combine data collection for the two surveys. Although the surveys have different questionnaires, collection and follow-up are done for both surveys in one telephone call. A further justification for this approach is the fact that the RCS can be regarded as a supplement to the MRTS. The latter gathers total monthly retail sales, while the RCS asks respondents for a breakdown of sales by commodity group. Since we were able to use the MRTS's infrastructure to collect RCS data, development of the RCS's collection system focused on development of the questionnaire, data capture system, edit rules specific to the RCS, and data transmission.

The RCS questionnaire lists over 100 commodity groups, which are themselves assembled into major commodity groups such as food, clothing and accessories, and furniture and appliances. The total sales for all major groups equals total retail sales. Respondents can report their sales by commodity group as a dollar amount or as a percentage of their total retail sales. If the respondent is unable to provide the data in either form, the interviewer will attempt to at least find out what types of commodities the respondent sells. This information can be input to the collection system and subsequently used at the edit and imputation stage to determine what fields need to be imputed. Since the majority of companies are in the survey sample month after month, we try to tailor each company's questionnaire to its industry and commodity groups, as reflected in previous responses. The tailored questionnaire eases the response burden and helps boost the response rate. The first time a unit is contacted for the RCS, the interviewer creates a profile containing a list of the commodities usually sold by the unit. The profile is used initially in preparing the tailored questionnaire and later in edit and imputation. It is updated regularly.

4.3 Editing and Imputation

As stated above, the RCS uses a complex questionnaire that includes many different commodity groups. These groups in turn form other groups, and so on. In summary, the questionnaire is filled with totals and subtotals and, as a result, it was difficult to develop an editing and imputation strategy. The strategy we devised was implemented within a tight schedule and was not tested since there were no usable historical data. Consequently, we tried to keep it simple, robust and flexible. The editing and imputation system consists of three main modules: pre-editing, automated editing, and imputation. Each module is described in detail below.

4.3.1 Pre-Editing

The purpose of pre-editing is to perform a series of checks on the data supplied by the units that contribute the most to the estimate of total sales in each retail trade sector. Those units may be either large companies or small companies that have a high sampling weight. The data they provide are checked to ensure that sums of parts and totals add up, that reported commodities match the type of business, and that there are no sudden changes in sales from month to month or year to year. Data that fail pre-editing are examined by subject-matter experts, who either correct the most obvious errors or contact the respondent for clarification.

4.3.2 Automated Editing

All data, even those which have undergone pre-editing, must go through the automated editing stage. The object of automated editing is to identify fields requiring imputation, while altering the data reported by respondents as little as possible. Automated editing finds erroneous data that must be replaced with imputed values. It allows the substitution of an imputed value for a reported value only if the reported value is involved in sums of parts that do not equal the totals. This is the only rule in the automated edit stage because without historical data we would have had great difficulty in setting the boundaries between the acceptance region and the rejection region for any other type of rule. Of course, this strategy may be reviewed after a few years of production. Checking sums of parts and totals may appear simple, but it is actually very complex because the subtotals are added together to form other totals. When it encounters non-response, the automated editing system tries to determine which of the fields involved should be zeroed and which should be imputed. The profile created during data collection, historical data (for the previous month and eventually the same month of the previous year) and even the unit's industrial classification are used in this process.

4.3.3 Imputation

Prior to actual imputation, a final series of checks is performed on the records that might be used to calculate values imputed to other records. The checks ensure that no outliers will be employed in those calculations. To that end we apply rules of the same type as those used in pre-editing, though the rejection and acceptance regions may be different. This step is designed to prevent situations such as the following: if a women's clothing store that also sells cosmetics were used, it might generate cosmetics sales for all non-respondent women's clothing stores.

The first step in the imputation process involves defining imputation groups. An imputation group consists of a set of homogeneous units. A value imputed to a unit will usually be derived from the values of respondents belonging to the same imputation group. In other words, we want to use units with similar profiles in the imputation process. The RCS's imputation groups are defined on the basis of the latest information about industrial classification, geographic area and unit size.

Ratio imputation and adjusted historical imputation are the methods currently used in the RCS. Let $y_{i,t}$ be unit i's sales of commodity Y at time t. Let $y_{i,t-1}$ and $y_{i,t-12}$ be unit i's sales of commodity Y the previous month and the previous year respectively. Finally, let $x_{i,t}$ be unit i's total sales at time t. If unit i's sales of commodity Y have to be imputed, we will use $y_{i,t}^{*}$, which can be calculated in one of the following ways.

Ratio Imputation:

$$y_{i,t}^{*} = \frac{\sum_{j \varepsilon G} y_{j,t}}{\sum_{j \varepsilon G} y_{j,t-12}} y_{i,t-12}, \tag{1}$$

$$y_{i,t}^{*} = \frac{\sum_{j \varepsilon G} y_{j,t}}{\sum_{j \varepsilon G} y_{j,t-1}} y_{i,t-1}, \tag{2}$$

$$y_{i,t}^{*} = \frac{\sum_{j \varepsilon G} y_{j,t}}{\sum_{j \varepsilon G} x_{j,t-1}} x_{t} \tag{3}$$

where G is the set of all units that responded to commodity Y in unit i's imputation group and passed the above-mentioned checks. Methods 1 and 2 ensure that imputed commodity Y sales have the same relationship to the previous year's or previous month's sales as those of respondent units in the same imputation group. Method 3 ensures that the percentage of imputed commodity Y sales in relation to total sales (X) is the mean percentage for responding units in the same imputation group.

Adjusted Historical Imputation:

$$y_{i,t}^{*} = \frac{x_{i,t}}{x_{i,t-12}} y_{i,t-12}, \tag{4}$$

$$y_{i,t}^{*} = \frac{x_{i,t}}{x_{i,t-1}} y_{i,t-1}. \tag{5}$$

Methods 4 and 5 ensure that the percentage of imputed commodity Y sales in relation to total sales (X) equals the percentage of commodity Y sales the previous year or the previous month.

Although edit and imputation are applied to the dollar values of commodity sales, the survey is much more concerned with the distribution of commodities, *i.e.*, the proportion of sales of each commodity in relation to total retail sales. Consequently, imputation methods 1 and 2 are better for the RCS than methods 4 and 5, and even method 3, since the latter methods tend to conceal changes in the distribution. In addition, wherever possible, imputation is performed within imputation groups defined by industrial classification, geographic area and company size. However, some groups do not contain enough respondent units, and we have to broaden the definition by removing geographic area, for example. Finally, since imputation does not ensure that the parts will add up to the totals, it is followed by a prorating step.

4.4 Estimation

First phase information has been heavily used in every stage of the RCS so far, and the next stage, estimation, will be no different. Since total sales are known in phase one, we chose an estimator that would enable us to use that information while maintaining a degree of simplicity in the estimation of variance without sacrificing precision. We had to explicitly develop a variance estimation formula for a two-phase design in which the second phase sample is selected from a restratified first phase sample. Two estimators were studied: the double-expansion estimator and the reweighted expansion estimator of Kott and Stukel (1997). For further details, see Binder *et al.* (1997). According to Jocelyn, Brodeur and Babyak (1997), a simulation of a double-expansion estimator produced results comparable to those of the reweighted expansion estimator. However, the double-expansion ratio estimator was selected because the variance estimator is very simple. The rest of this section deals exclusively with the double-expansion ratio estimator. Let $\hat{T}_k$ be the estimator used to measure the total sales of commodity *k*. It can be written as follows:

$$\hat{T}_k = \frac{\hat{Y}_k}{\hat{X}_k} \hat{X}_k^{(1)}$$

$$\hat{X}_k^{(1)} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j=1}^{n_h} x_{hjk}$$

where $\hat{X}_k^{(1)}$ represents the estimated sales based on the first phase sample (MRTS) $\hat{Y}_k$ represents the estimated total sales of commodity *k*, and $\hat{X}_k$ represents the estimated sales from the second phase sample.

The variance estimator of this estimator is given by:

$$\hat{V}(\hat{T}_k) = \hat{V}_1(\hat{T}_k) + \hat{V}_2(\hat{T}_k)$$

with $\hat{V}_1(\hat{T}_k) = \sum_{g=1}^{G} M_g^2 (1 - f_g^{(2)}) \frac{s_{2g}^2}{m_g}$

and

$$\hat{V}_2(\hat{T}_k) =$$

$$\sum_{h=1}^{H} \sum_{g=1}^{G} \frac{N_h^2 (1 - f_h) M_g^2 (1 - f_g^{(2)})}{n_h^2 (n_h - 1)} \frac{s_{1(h)g}^2}{m_g} +$$

$$\sum_{h=1}^{H} N_h^2 (1 - f_h) \frac{s_h^2}{n_h}$$

where $s_{2g}^2, s_{1(h)i}^2, s_h^2$ are defined in Binder *et al.* (1997), and $f_g^{(2)} = m_g / M_g f_h = n_h / N_h$.

This variance estimator was developed with the Taylor linearization method (Binder (1996)).

5. CONCLUSION

The RCS was developed very quickly. Choosing a restratified sampling plan for the second phase helped us keep to a very tight schedule; it also improved the design's efficiency and reduced costs by using auxiliary information. At present, the data are being collected monthly, but the results are published quarterly. Hence it is essential to estimate the covariance since the samples are not independent from month to month.

A few challenges remain. For example, it will be beneficial to track commodity imputation over time and adjust the imputation strategy as required. Also, the commodity data for many companies remain quite stable from quarter to quarter. We are considering switching to annual collection in those cases.

6. ACKNOWLEDGMENTS

The authors wish to thank all those who worked on the methodology of the RCS, including Colin Babyak, Hélène Bérard, Sarah Franklin, Julie Girard, Leslie Kiss, Martin St-Pierre and Janet Sear. The authors would also like to thank François Maranda and Don Royce for their invaluable comments.

REFERENCES

Binder, D.A. (1996). Linearization methods for single phase and two phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M., and Jocelyn, W. (1997). Variance estimation for two-phase stratified sampling. *Proceedings of the Section on Survey Research Methods*, Annual American Statistical Association. To be published.

Cochran, W.G. (1963). *Sampling Techniques*. John Wiley.

Hidiroglou, M.A., and Särndal, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods*, Annual American Statistical Association, 873-878.

Jocelyn, W., and Brodeur, M. (1996). Méthodes de répartition multivariées pour l'échantillonnage à deux phases: Application à l'enquête trimestrielle sur les marchandises. *Recueil des communications des XXVIIIe Journées de Statistiques de l'ASU*, 433-436.

Jocelyn, W., Brodeur, M., and Babyak, C. (1997). Comparaisons de différents estimateurs de variance à deux phases: étude Monte-Carlo basée sur l'Enquête des marchandises au détail. *Recueil des de la section des méthodes d'enquête*, SSC, juin 1997.

Kott, P.S., and Stukel, D.M. (1997). Can the Jackknife be used with a two-phase sample? *Survey Methodology*, 23, 81-89.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Sampling. Springer-Verlag.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

THE WORKPLACE AND EMPLOYEE SURVEY: METHODOLOGICAL CHALLENGES AND PILOT RESULTS

H. Krebs, Z. Patak and T. Wannell[1]

ABSTRACT

The Workplace and Employee Survey (WES) is a new undertaking by Statistics Canada with funding from Human Resources Development Canada. WES is a dual survey that starts with a sample of establishments and then draws a sample of employees within those establishments. Employers and employees are administered separate questionnaires covering a broad range of workplace issues. Thus information from both the supply and demand side of the labour market will be available to enrich studies focused on either side of the market. A large-scale pilot of WES was carried out in the winter and spring of 1996. This paper focuses on the challenges identified in a large-scale WES pilot and the plans for a production scale survey.

KEY WORDS: Workplace survey; employee survey; Business survey; Longitudinal survey.

1. INTRODUCTION

The nineties have witnessed a bumper crop of buzzwords for anyone interested in the economy in general and the labour market in particular. We are working in a new competitive environment, making the transition to a knowledge-based economy. To thrive in this environment, firms must be flexible or adaptive; they should develop high performance workplace practices. Employees too must be adaptive; they can empower themselves by adjusting their skill set. Otherwise they risk becoming disposable.

Annoying as they may be, clichés don't reach that status without some underlying truth. Canadian firms and their employees have always faced a competitive, changing environment. Some types of change – particularly those related to microprocessor technologies – have probably quickened pace in recent years. The development of a North American free trade zone has certainly heightened awareness of the competitive environment. And the growing disparity among workers (and would be workers) – both in terms of earnings and hours – has been well documented. These trends contribute to a general sense that economic change is increasingly difficult to understand, that the costs of change fall mainly upon less-adaptable workers and that even among the "winners" in the labour market, employment is becoming less stable.

Looking at these and other problems, analysts in Statistics Canada and elsewhere have reached the conclusion that there are two key elements missing in our understanding of firm performance and worker outcomes. The determinants of how well firms respond to change can only be properly studied in a longitudinal setting that covers many of the firm characteristics and behaviours related to performance. Of particular importance are the practices and policies related to employees, since they must be the agents of change in the firm. Conversely, the fortunes of employees are intricately tied to what they do on the job and how they interact with the internal forces of change in a firm. Thus the ideal survey instrument would follow an integrated sample of employers and employees over an indefinite period. Some of these elements exist in other Statistics Canada surveys, but not in an integrated design.

Recognizing the importance of dynamics in general, Statistics Canada has steadily increased its capacity to follow businesses and individuals longitudinally. Both survey and administrative data have been linked longitudinally to follow the fortunes of firms and workers. Household longitudinal surveys, such as the Labour Market Activity Survey and the new Survey of Labour and Income Dynamics, have job mobility and earnings dynamics as primary concerns. These and other data sources led to new insights on cyclical patterns in quits and layoffs, movements into and out of low income status and the repeat use of the unemployment insurance system, among others.

At the same time, clients interested in the competitive position of Canadian industry have sponsored surveys on technology use, innovation and the success of small and medium sized enterprises. These surveys are beginning to shed some light on firm growth and decline, particularly how the adaptive and innovative capacities of firms contribute to their success.

Despite this recent activity, fundamental information gaps pertaining to the workplace remain. We now know a lot about the magnitude of hiring and separations, but, particularly in the case of hiring, not much about what triggers this activity. We know very little about actual and perceived job stability. Our information on performance-based and other non-standard types of pay is very patchy. We know which firms use particular technologies, but little about which employees use them and how it affects their skill requirements and pay. Our knowledge of contract (or contingent or temporary) workers is very limited. We

[1] Ted Wannell, 24-F R.H. Coats Building, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; e-mail: wannell@statcan.ca.

cannot adequately construct earnings distributions within firms. We know almost nothing about how employee compensation and other human resource practices relate to firm performance. We obtain information on the total training offered by firms, but know little about who in the firm receives it, how much they receive or how training relates to worker turnover, firm performance and employee compensation.

These gaps hamper the development of policy pertaining to both human resources and industrial strategies. Although these policy areas are closely related in reality, the data resources and hence the research on which policy is based, have tended to be isolated. This is in part because at the level of data resources, information on firms and their workers has not been integrated. The inability to link information on (1) activities within establishments influencing labour demand and worker outcomes, and (2) the characteristics of the workforce, has meant that research on both establishments and labour market issues has suffered.

The Workplace and Employee Survey (WES) is a new Statistics Canada undertaking, sponsored by Human Resources Development Canada, that is designed to provide an integrated view of the activities of employers and their employees. A large-scale pilot of the WES was conducted in 1996, with a production survey scheduled for 1998. In the remainder of this paper, we provide an overview of the WES pilot study, look at some of the methodological problems and discuss future plans.

2. WES PILOT SURVEY OVERVIEW

To test both the feasibility and efficacy of a dual survey to address some of the issues noted in the introduction, both pre-testing and a large-scale pilot were conducted. Early pre-testing confirmed that employers were able to answer the type of questions proposed and provide lists of employees from which intra-establishment samples could be drawn. Human Resources Development Canada provided funding for a large-scale pilot to test more fully the operational, methodological and analytical feasibility of the project.

The pilot aimed to interview approximately 1,000 employers in selected strata from a production-scale sample of 5,500 employers. Up to seven employees would then be sampled within each selected establishment. The remainder of this section outlines the content of each of the surveys, sketches the frame creation and sampling methodology and summarizes the operations.

2.1 Survey Content

Two separate questionnaires were developed for the pilot: one for employers and one for employees. The employer questionnaire contains a broad range of information. So broad, in fact, that we anticipated that several respondents may be required to answer it completely, particularly in large establishments. As such, the questionnaire was parceled into blocks – each with a separate cover sheet – so that each block could be directed to the appropriate respondent. A brief description of each block follows.

- **Workforce Characteristics and Job Organization:** Covers the work arrangements of employees (full-time/part- time, permanent, seasonal, on-site/off-site, *etc.*), recent hiring and separations, and the presence of unfilled vacancies. All questions in this section were broken down into five occupational groups.
- **Compensation:** Covers variable pay plans, gross payroll, non-wage benefits and the distribution of earnings in the company. Most questions captured occupational detail.
- **Training:** Covers the presence of formal training programs, which occupational groups received training in the past year, how training was funded and how much was spent on training.
- **Human Resource Function:** Determines who has responsibility for human resources, the level of employee involvement in decision-making, and the incidence, type, extent and effects of recent organizational change.
- **Collective Bargaining:** Asks about the presence and membership (by occupation) of collective bargaining groups, treatment of "flexibility" issues in contracts, work stoppages and grievances.
- **Establishment Performance:** Covers operating revenues and expenditures, change from the previous year, variability in revenues by quarter and foreign ownership.
- **Business Strategy:** Asks respondent to rate the importance of elements of business strategy, estimate their distribution of sales by market area and specify the number of competitors in their market.
- **Innovation:** Identifies major innovations introduced in the past three years.
- **Technology Use:** Asks about overall computer usage in establishment, looks at specific major technology implementations in the past three years (hardware/ software, computer-controlled technologies and other technologies) and the effects of the implementations.
- **Use of Government Programs:** Looks at establishment use of grants and loan, employee-related programs, tax provisions, information services and other ventures with government.

The employee questionnaire was not as clearly blocked as the employer questionnaire, since it only involved a single respondent. The questionnaire covered: job characteristics, requirements when hired, hours of work, pay and benefits, working off-site, leave, promotions, technology, training, participation in decision-making, work stoppages, recent work history, education, family situation and membership in designated employment equity groups. While the questionnaire covers a fairly wide range of topics, the pilot demonstrated that it was not overly burdensome for respondents[2].

2.2 Survey Frame and Sampling

The WES is based on the notion of a workplace as the microdata unit where labour supply and demand is resolved. Although the responsibility for staffing is included in this

[2] Typically, interviews lasted about 25 minutes.

concept, it more importantly includes the organization of a group of employees to achieve a common purpose. Our ultimate target population includes workplaces in all industries and geographic areas of the country. Ideally, WES would operate as a two-stage survey. The first stage would involve drawing a sample of workplaces that is large enough to produce estimates for homogenous industries at the provincial level. The second stage would draw a large enough sample of workers within each workplace to permit variance calculations. In conducting a survey, however, our concepts and intentions are tempered by operational constraints and the availability of data.

Statistics Canada's Business Register (BR) – a registry of all businesses in Canada – is the primary frame resource for business surveys. The BR organizes business entities into a hierarchy of four statistical levels: enterprises, companies, establishments and locations. Although the location level is conceptually the closest to a workplace, several factors led us to sample from the establishment level[3]. An establishment can be thought of as the smallest organizational unit, comprised of at least one physical location, that can provide a complete set of input and output statistics. For most businesses, establishments and locations are one and the same. However, establishments in many larger enterprises – particularly those in the financial, communications and utilities sectors – may include separately managed operations in a number of locations.[4] For these complex units, WES sampled smaller units within the establishment using information from the BR, auxiliary files and, in rare cases, from contact with respondents. Thus the employer survey evolved into a two-stage sample and the employee survey a three-stage sample.

At the first stage of sampling, the frame is stratified by region, industry and employment size. Sampling fractions vary by size group so that larger employers have a greater probability of being included in the sample. In the second stage, complex establishments drawn in the first stage are subdivided into smaller units and a sample of these units is drawn. In the third stage, a sample of workers is drawn from employer-provided lists in each workplace.

2.3 Operations

The unique content and methodology of WES placed unusual demands on survey operations. Many of the required operations had no recent precedents at Statistics Canada. What we provide here is a thumbnail sketch of the survey operations without too much detail on the logistical permutations involved.

Preparation for fieldwork began with the examination of the sample (primary sampling units – PSUs) for potentially complex establishments. Survey staff used the Business Register and SEPH to determine whether there were any known operational units (secondary sampling units – SSUs) below sampled establishments. SSUs were selected at a rate of one per stratum – given that some stratification (*e.g.*, by size or function) was evident among the SSUs. The employment of all SSUs was recorded and forwarded to survey methodologists for estimation purposes. The coverages of selected SSUs were forwarded to operations.

Headquarters operations developed interviewer training and resource materials and provided the training to field staff in each region. They also directed the sample of PSUs and SSUs to the appropriate regional office. Interviewers in the regions contacted employer respondents to schedule on-site interviews. Interviewers had a number tasks to perform in these visits.

- Complete all possible sections of the employer questionnaire with the available respondents. Document any problems regarding survey content or procedures.
- Leave appropriate sections of the questionnaire behind when required respondents are not available.
- Take sample of employees from employer list according to methodology's instructions. Record names of and, at interviewer's discretion, other information about sampled employees.
- Leave employee information/consent packages with the primary respondent to distribute.
- Transmit sampled employees' names to Winnipeg RO.

Sampled employees were to fill out the contact/consent forms – which asked for information on convenient times and numbers for a telephone interview – and return them by fax or pre-paid mail. About two-thirds returned the forms and about 85 percent of those agreed to participate. Employees were interviewed by phone from the Winnipeg RO. All the information in the employer and employee interviews was recorded by pencil and paper and later captured in a Microsoft Access application.

The survey was in the field from December 1995 to April 1996. Data capture and preliminary validation were completed by mid-July 1996. After accounting for dead establishments and non-employer businesses (which were deemed out of scope), the response rate for the establishment survey was approximately 80 percent. The response for the employee survey was 56 percent of the employees sampled, somewhat lower than that if one takes into account the truncation of the sample due to establishment non-response.

3. METHODOLOGICAL PROBLEMS AND RESPONSES

3.1 Primary and Secondary Sampling Units

The employer portion of WES was originally conceived as a stratified single stage design with establishment as the primary sampling unit (PSU). The population was partitioned by industry (14 categories), region (6 categories), and size (industry and region dependent) from which a simple random sample was selected without replacement. It became apparent that, for approximately ten percent of the sample, the target unit of interest – Workplace Reporting Unit (WRU = SSU = secondary sampling unit) – corresponding to a physical location, was different from

[3] For more details, see "Frame Allocation for the Workplace and Employee Survey", Sharon Wirth, Business Surveys Methods Division.

[4] The sub-establishment units sampled were not necessarily statistical locations. Please see the section on Workplace Reporting Units for more details.

the PSU. This came about as a result of many larger establishments having multiple workplaces (locations on the Business Register).

Collecting data from every location of a complex establishment was not feasible due to sample size constraints imposed on the pilot survey. A second stage was added to the survey design to facilitate the sub-sampling of PSUs. Each in-sample complex establishment was stratified by type of WRU (*e.g.*, Head Office, typical bank branch, *etc.*). For the pilot one location was selected from each WRU stratum. Sampling at a higher rate at the second stage was deemed fiscally and operationally prohibitive.

The employee portion of WES added a second/third stage to the employer survey. After a WRU had been sampled, a list of employees was obtained from the employer followed by the selection of a systematic sample of six (or seven) employees. WRUs with fewer than seven employees were sampled exhaustively. The pilot sample consisted of approximately 3,500 employees, of which 1960 responded to the survey, representing 544 WRUs.

3.2 Multi-Stage Estimation

In a typical multi-stage survey the estimated variance is the sum of estimated variances from each stage, provided that at least two units have been selected in each stratum at each stage. Failing to satisfy this criterion (WES sampled one unit per stratum in the second stage of selection) forced us to find an alternative to estimating proper multi-stage variances. To that end, we made the simplifying assumption that the first stage units had been selected with replacement and proceeded to compute the corresponding variances.

The Statistics Canada Generalized Estimation System (GES) was used to compute the second stage weights for multi-location establishments. The sample weight was set to unity for single-location establishments. The stratification of SSU's was ignored for estimation. Ratio estimation was used with employment as the auxiliary variable. The total establishment employment, which was not collected, was derived form the Business Register. The BR employment values were adjusted by a multiplicative constant so that the adjusted BR numbers would add up to SEPH (Survey of Employment, Payroll and Hours) estimates.

A second pass through GES produced first-stage weights for the sampled establishments. Again, ratio estimation was used with employment as the auxiliary variable. Control totals were applied at the industry/region level with an exception discussed below.

GES was also used to compute sample weights for employee records. An employee was given a weight, ignoring non-response, equal to the number of employees in an SSU over the number of sampled employees. This, in fact, was a calibrated weight, since the number of employees in the SSU was taken from the employer questionnaire. The reported SSU employment could, conceivably, be different from the number of employees on the list used by the interviewers for sampling. Unfortunately, the number of employees on the employer lists was not recorded; it could have provided a measure of non-sampling error.

3.3 Harmonization of Employer and Employee Estimates

Auxiliary information used for ratio estimation at the first stage was available at the cell level (industry/ region/ size). To improve the stability of the employer and employee estimates, and to ensure consistency between the two portions of the survey, auxiliary totals needed to be rolled up to the industry/region. An exception was made for industry 11 – Finance and Insurance – where two control totals were used: one for Quebec and another one for the rest of Canada. This was precipitated by a relatively high number of establishments with no employee data. In fact, some employee estimation cells at the lowest level of stratification were empty. Note that total employment by industry/region was obtained from SEPH.

3.4 Pilot Meta Results

The employer sample consisted of 1,006 live, 53 dead, 54 inactive, 1 receivership, 11 holding company and 169 out-of-scope PSUs. Estimates of totals for some 897 variables were computed using 1025 establishments (all except "live/complete refusal"). At the national level, the coefficients of variation (CV) for Gross Operating Revenue, Gross Expenditures and Total Gross Payroll were 0.0887, 0.0654 and 0.0201 indicating good reliability. Overall, still at the national level, two thirds of the estimates had a CV between 0 and 0.33.

On the employee side 1960 persons provided either partial or complete response. As an example of reliability of the totals computed from the employee portion, the CVs, at the national level, for Family Income and Salary were 0.0236 and 0.0230. Overall, still at the national level, three quarters of the estimates had a CV between 0 and 0.33.

4. FUTURE PLANS

The value of the WES will be enhanced as its workplace panel is tracked over time. For example, cross-sectional studies of the correlates of workplace success (or lack thereof) are hampered by two major problems: the direction of causation and survival bias. These methodological issues can be overcome by true panel data. While the pilot WES is a cross-sectional survey with retrospective questions, current plans are to develop an annual longitudinal production survey for employers and their workers. Such a goal will aid policy-makers and researchers in particular in understanding the reasons and trends of changes taking place as employers and their employees act and interact within the workplace. The intention is to design the survey to address issues of interest that require both cross-sectional and longitudinal data on employers and their employees.

The following topics are merely a few of the enormous range of research topics that could be explored with longitudinal data.

a) The stability of today's jobs and the emergence of "core" and "periphery" workforces. The longitudinal aspect of the survey will document changing employment arrangements into the future.

b) The causes and characteristics of successfully growing firms in order to secure or create jobs. The WES will enable the tracking of changes in employment and its relationship to firm performance.
c) The relationship between training, technology use and job stability. To study if employees who receive training and/or use technology intensely are less likely to change jobs or experience unemployment requires at least two periods of data to calculate transition rates.
d) The influence of technology and international trade on wage distributions. Longitudinal data about the changes in the distribution of wages in firms that, for example, have adopted technology and/or function in competitive markets would provide a means for testing if these mutually exclusive phenomenon drive the polarization in wages and labour demand observed in the labour market.

4.1 Basic Elements of the WES Longitudinal Design

4.1.a Workplace

Our working assumption is to follow workplaces for a long time. Businesses determined to be out-of-scope in the first production survey will be monitored before the second survey occasion for status changes. This will be done using a combination of Payroll Deduction account and Business Register frame information. If the workplace status has changed to in-scope in the second year, then the workplace will be included in the sample of "continuing" workplaces. Interviewers will also make every effort to convert previous year's non-responding workplaces.

In the third year and in each subsequent odd year sample attrition will be handled by selecting new entrants from a pool of births using the Business Register frame. Changes over time in the statistical universe are a concern for longitudinal surveys. There are several circumstances under which a member of the WES panel could continue in business, yet be registered as a death and subsequent birth – with a new identifier – on the Business Register. Such events necessitate adjustments to the cross-sectional weights.

4.1.b Employee

An "ideal" employer-employee survey would follow employees for long periods of time, as well as employers. Employees, however, may change employers with some frequency. To follow them from employer to employer would result in an explosion of our employer sample size. Accordingly, our original plan was to follow employees for as long as they were with the establishment they were originally sampled in and for one period thereafter. Similar to the establishment sample, sample attrition would be handled by selecting new employees from a pool of new hires since the first sample. However, a follow-up to our pilot showed that not all employers could put together lists of new employees. Rather than continue with an asymmetrical sample, a different strategy was developed.

Employees sampled in the first year will be interviewed for two consecutive years. An "exit" questionnaire will be administered to those employees no longer with the same employer in the second year. The employee sample will be replenished in the third year from a pool of all current employees yielding a new cohort of employees. The two-year cycle would then be repeated for the selected employees in the third year. A set of retrospective questions for recently hired employees was added to the first cycle questionnaire, so that transitions into and out of the employer would be covered.

4.2 Longitudinal Design Issues

The WES plan to follow both establishments and employees in a single vehicle is new to Statistics Canada. On the household side, the Survey of Labour and Income Dynamics tracks individuals over time, along with the changes in their family situation. Methodological work has been focused on developing and maintaining cross-sectional and longitudinal weights for both units. In the WES, we will be following cohorts of establishments over long periods of time, along with shorter panels of employees within those establishments. Our goal too is to create and maintain both cross-sectional and longitudinal weights, but the challenges are somewhat different.

A research paper summarizing the pilot survey changes necessary to implement the longitudinal tracking of employers and employees for the production survey will be available in April 1998. Changes related to the following issues will be addressed:

- rules regarding the tracking of workplaces and their employees;
- workplace replenishment;
- turnover of workplace contacts over time (conditional interviewing);
- weighting strategies for period-to-period transitions and changes in employee earnings;
- expected turnover rates for employees and the replenishment of employees through time;
- attrition due to non-response;
- changes in ownership and organization of workplaces; and,
- strategies to track exiting workers.

4.3 Inaugural Production Survey

The initial production survey is scheduled to be in the field in April 1998. The planned usable sample for this survey will be about 6,000 workplaces and up to 30,000 employees. The target population for the Workplace component is all employers operating in Canada having paid employees, with the following exceptions:

a) employers in the Yukon and Northwest Territories
b) employers operating in agriculture and related services; fishing and trapping; highway, street and bridge maintenance; federal, provincial and local government services; international and other extra-territorial government services; private household and religious organizations.

The target population for the Employee component is all employees working in the selected workplaces who receive a Revenue Canada T-4 Supplementary form. In order to improve coverage of occupations and reduce overall sampling variance, up to 12 workers will be selected in the larger workplaces.

The population will be stratified into 14 industry groups based on the 1997 North American Industrial Classification Standard (NAICS). The geographic breakdown will be the Atlantic region, Quebec, Ontario, the Prairies (Manitoba and Saskatchewan) Alberta and British Columbia.

The 1991 Standard Occupational Classification will be used to post-code employees' job titles and work activities from the Employee survey portion. The mapping of aggregate level occupations between the workplace and employee survey portions will permit the validation of the consistency of responses between employers and employees for questions with similar subject matter. This analysis can determine if knowledge and awareness of programs, benefits and policies differ by workplace or employee characteristics.

The survey reference period for the Workplace portion will mainly be the 12-month period ending March 1998. For questions on employment by occupation detail, compensation practices, hours of work, workers covered by collective agreements and general business strategies, the reference period will be the last pay period in March 1998.

The collection window for the Workplace and Employee portions will be the April-August 1998 period. The plan is to process the survey data over the May 1998-August 1999 period culminating with the availability of "clean" employer and employee microdata files in September 1999. The current plan is to explore the possibility of releasing a micro data file for the Employee portion only. A publication summarizing the basic results of the initial production survey is planned to be released by November 1999.

Addendum: WES data collection will most likely be delayed by one year from the dates mentioned above due to the phasing-in of funding for the project.

SESSION I-2

Use of Administrative Data in Censuses

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

STATISTICAL PROPERTIES AND QUALITY OF REGISTER-BASED CENSUS STATISTICS IN FINLAND

R. Harala[1]

ABSTRACT

The use of administrative data for statistical purposes has increased in Finland during the last twenty-five years. A good example of the trend are the population censuses in which the use of register data has increased census by census. The population and housing census of 1990 was the first one to be conducted totally register-based without any questionnaires to the population. In the census system over thirty registers are used. The most important advantages of using administrative data in censuses and in the production of statistics in general are the savings in the costs of data collection, the reduction of the response burden of the population and enterprises and as a result of these, the possibility to produce data more often. Actually nowadays the population and housing census type data are produced annually in Finland.

KEY WORDS: Register-based population census; Administrative records; Register estimation.

1. BACKGROUND OF USING ADMINISTRATIVE DATA FOR STATISTICAL PURPOSES

There has been a continuous increase in the use of data from administrative records for statistical purposes during the past twenty – twenty-five years in Finland and also in other Nordic countries. Actually today about 94% of the data collected by Statistics Finland comes from administrative sources and only 6% of the data collection is based on direct data collection from population or enterprises.

There are certain factors that have facilitated the use of administrative registers in Finland the most important ones of which are:

Use of unique identification numbers. In Finland a unique personal identification number is used and we also have some other almost as uniform identification number systems (identification of enterprises, houses, dwellings) which are used in all central registers, above all in those that involve matters of social security and taxation.

Administrations own interest in building nation-wide databases. An important factor has also been the fact that most systems on social security and taxation are on a national level or that administration itself has had an interest in combining them into one nation-wide database. Furthermore in recent years more efforts have been made to centralize the distribution of the basic allowances of social security to one and the same authorities. The majority of Finnish social benefits are also taxed which means that the taxation authorities receive the information from the social security institution for taxation.

Acceptance of the population. A very important factor facilitating the use of registers has also been the fact that there has been a wide acceptance of the population of the use of data collected for administrative purposes also for statistical purposes. Populations trust in society and the rationality of its functions has been rather strong. Also the reputation of Statistics Finland in matters of confidentiality is very good. The acceptance comes from the fact that people consider it rational to use data already collected for administrative purposes for statistical purposes too, because it is much cheaper than to collect the same data on questionnaires. It also reduces the response burden of people and enterprises which is also a very important reason of using data already gathered somewhere.

Legal basis. The legal basis of using administrative data for statistical purposes is of course also a very important factor. The legal basis of using administrative registers in Finland is a reflection of the widespread opinion that it is rational to use data already in the possession of some administrative authority and that is why our Statistics Act is based on the principle that administrative data should always be used when they are available. Statistics Act also grants Statistics Finland wide possibilities to get, to use and to link administrative records. It also defines the data confidentiality which we think is a very important issue in the statistical legislation.

2. USE OF ADMINISTRATIVE DATA IN POPULATION CENSUSES

The best example of the trend of using more and more data collected for administrative purposes also for compiling statistics is the population and housing census system in Finland. Finland has long traditions in using administrative records in the production of census statistics. The use of register data has increased census by census after the census of 1970 in which administrative data were first used.

The 1990 population and housing census was for the first time in Finland conducted entirely on the basis of registers

[1] Riitta Harala, Statistics Finland, FIN-00022 Statistics Finland, P.O. Box 4A, FIN - 00022; e-mail: riitta.harala@stat.fi.

without a questionnaire addressed to the whole population. For the time being the 1995 census is carried out on the basis of the same system and we are going to use the same method also in the census of 2001.

The construction of a register-based census system began already at the beginning of the 1980s primarily for *reasons of cost*. The *administrative registers were by that time sufficiently developed* to make a register-based system possible. The *growing need for statistics* was also one of the main reasons for developing a register-based system; the production of statistics by means of a register-based system is far more economical than by means of a questionnaire, which means a census can be carried out more frequently than once every five or ten years.

Actually this work has also resulted in producing census type statistics annually. The annual production of register-based building and housing statistics and the production of register-based statistics on the economic activities of the population (regional employment statistics) were both started in 1987. The users of for example annual employment statistics by municipality or by sub-region of municipality cannot even think the situation of not having these data every year. Regional employment statistics system has also become more and more important background data file in most of the new statistics systems of the population statistics unit in Statistics Finland. Examples are different kinds of flow statistics, educational indicators, statistics about foreigners *etc*.

2.1 Method

There are a variety of means available for using administrative records in compiling statistics. The ways of utilising administrative data can in principle be divided into the following categories, for example:

(1) direct tabulation,
(2) sampling and calculation frameworks,
(3) model-based estimation, and
(4) register estimation.

Statistics Finland uses all these methods in the use of registers. Register estimation is the primary method used in Finland in the register-based census system and its subsystem of regional employment statistics. It involves using questionnaire-based data to construct a model or an optimal group of decision rules by which one can derive from one or more items of administrative data a maximally good parameter value that conforms to statistical concepts to represent each unit of the population.

The method thus makes use of a number of sets of register-based data at the same time, so that deducing the main economic activity of individual persons, for example, requires the use of more than 20 sets of register data. This joint use of registers means that each register does not necessarily have to be fully comprehensive, as they can be used to supplement each other. Information on the same person may thus be available from different data sets, among which the most reliable information from the point of view of the material as a whole can be selected.

It should also be noted that the method is not especially vulnerable to changes in register data, as data on individual phenomena can usually be derived from a number of sources. This means that personal employment data are available via employment pensions systems, taxation registers and reports submitted by the individuals themselves or their employers. Should an employer neglect his obligation to inform the taxation authorities, for example, data will in any case be available through the pensions system or the individual's own tax returns.

2.2 Registers

The Finnish population census system draws on the basic registers maintained by society giving data on the population census target units and the links between them. These units comprise individuals, families, households, buildings (and summer cottages), dwellings (and business premises), enterprises and establishments (Annex 1).

The Central Population Register is a basic register in register-based census system, defining the population included. The population contained in the employment statistics comprises all persons whose legal domicile on December 31 is Finland. All the basic demographic characteristics of the population (identification number, sex, age, marital status, place of birth *etc*.) are derived from the Central Population Register.

The Register of Buildings and Dwellings is a basic register defining all buildings and dwellings. It also contains the characteristics of buildings and dwellings. The basic data of the register were collected in the 1980 population and housing census and they have been maintained since by Population Register Center in which the register is situated.

The Statistics Finlands Register of Enterprises and Establishments, which is not an administrative register, but which has properties of the same kind. For example branch of industry, ownership type, juridical form and location of the workplace are characteristics which are linked to employed from the Register of Enterprises and Establishments and from the Register of Public Associations and Local Government Functional Units.

Besides these basic registers the census system also uses about 30 other register (Annex 2) files, from which data are combined with the census units. These files include:

The Ministry of Labour's Register of Job Applicants

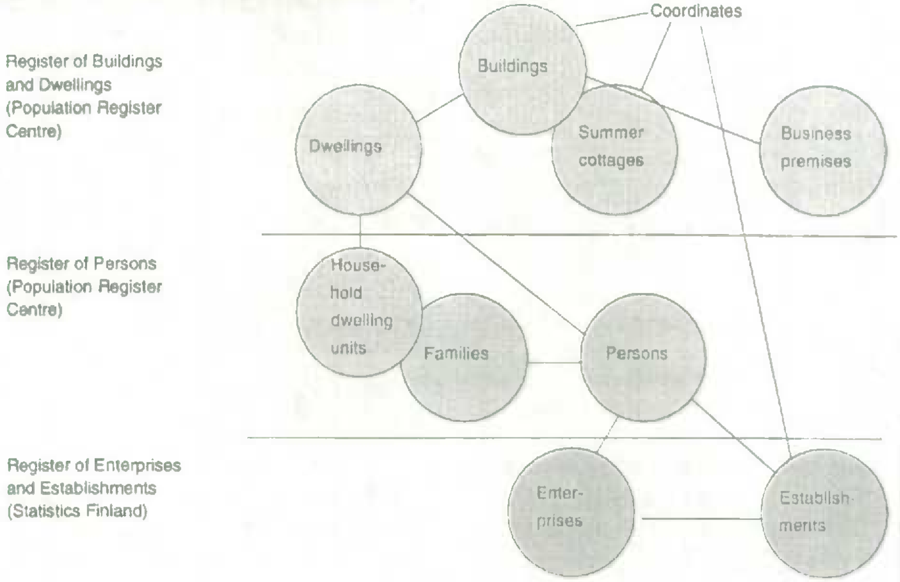
This register covers all persons who have applied for work through the employment exchange. The most important data derived from this register are:

- date when unemployment started and ended (each unemployment period of the year),
- reason for end of unemployment,
- date when job placement started and ended.

Information of conscripts and conscientious objectors are obtained from the General Headquarters Register of Conscripts and the Ministry of Labour's Register of Conscientious Objectors. The data obtained from these registers are personal identification number and date when military service/non-military service started and ended.

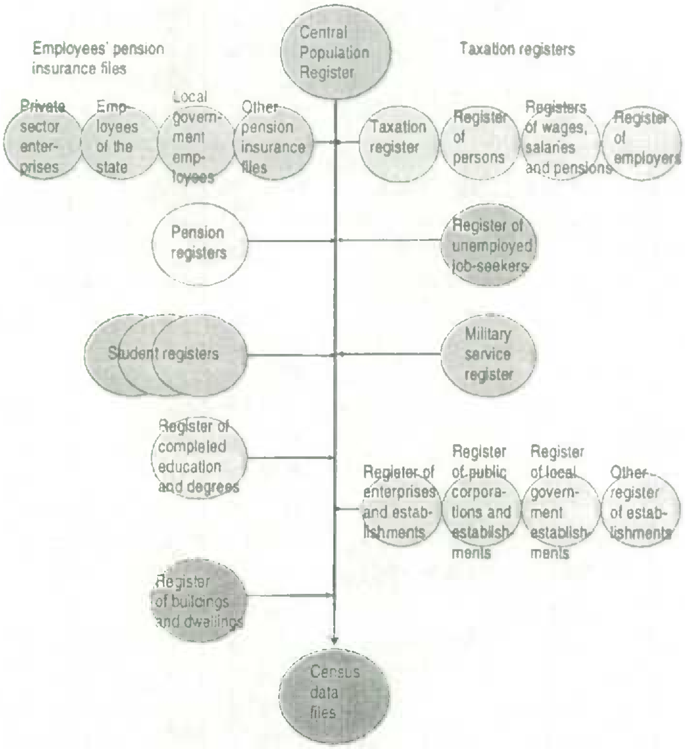
Annex 1

Units belonging to the register based census and links between them


Coordinates
Register of Buildings and Dwellings (Population Register Centre)
Buildings
Dwellings
Summer cottages
Business premises
Register of Persons (Population Register Centre)
Household dwelling units
Families
Persons
Register of Enterprises and Establishments (Statistics Finland)
Enterprises
Establishments

Annex 2

Use of registers and administrative records in the register-based census system


Employees' pension insurance files
Central Population Register
Taxation registers
Private sector enterprises
Employees of the state
Local government employees
Other pension insurance files
Taxation register
Register of persons
Registers of wages, salaries and pensions
Register of employers
Pension registers
Register of unemployed job-seekers
Student registers
Military service register
Register of completed education and degrees
Register of enterprises and establishments
Register of public corporations and establishments
Register of local government establishments
Other register of establishments
Register of buildings and dwellings
Census data files

The data on employment relationships are received from several different sources. These relationships can be divided into three different main groups: *employment relationships within the private sector (including the self-employed), within the government sector and within the local government sector (including municipal federations).*

The Central Pension Security Institute provides almost all the data on employment relationships within the private sector. It records all employment relationships covering a period of over one month during a year. In addition, it contains the dates when an employment relationship started and ended.

The Central Pension Security Institute provides data of which the following data are the most important:

- name of the enterprise,
- date when the employment relationship started and ended,
- type of pension scheme,
- pension regulation number.

Other sources of employment relationship data within the private sector are the National Church Board's Register of Employees, the Register of Seamen's Pension Scheme (from the Central Pension Security Institute).

The major part of the data on employment relationships within *the Central Government sector* comes from the State Treasury register of central government employees. These data comprise for example:

- office number,
- number of establishment,
- date when the employment relationship started and ended,
- occupation.

Other sources of data on employment relationships within the Central Government sector are the registers of the Social Insurance Institution, the Bank of Finland, Helsinki University and the Post and Telecommunications Office.

The majority of the data on employment relationships within the local government sector comes from *the Municipal Pension Institution*, which takes care of the register of the employment relationships of the local government and municipal federations. The data received are for example:

- member association number,
- date when the employment relationship started and ended,
- occupation.

The Employment Relationship Register of *the Government of Åland* is another source of information on employment relationships within the local government sector.

In connection with the work of collecting all employment relationships from different sources, a separate *employment relationship register* is compiled. This register comprises all the employment relationships during the year.

Data on wages/salaries for all employment relationships are obtained from *the Income and Pension Register of the National Board of Taxes* (Employer's statement of earnings and tax deductions). These data are needed when choosing the employment relationship valid at the end of the year.

The data on students are also obtained from a number of sources. *Statistics Finland's Register of University Students* provides information on persons studying at a university during the autumn term or registered at a university, as well as the type of university studies.

The Register of Study Aid to Students supplies data on study aid granted during the year (during the autumn and/or spring term).

The Joint Selection Register provides information about the students who have started their studies at the senior level in an educational institution during the year in course or the two previous years.

The Ministry of Education's *Register of Students in Senior Secondary Schools and Vocational Schools* supplies data on codes and locations of educational institutions of the students in these schools.

The Social Insurance Institution's Register supplies information on pensioners. The data used in employment statistics are the initiation date and type of pension.

The Register of the National Board of Taxation provides many different kinds of data on income and benefits as well as taxes, assets and liabilities of the persons subject to taxes.

Statistics Finland's *Register of Completed Education and Degrees* contains all the degrees taken in Finland as well as the majority of the degrees taken by Finnish citizens abroad.

The Register of Completed Education and Degrees provides data on the highest degree for the employment statistics. Since this register contains all degrees that a person has taken, it is possible to use these too. The information describing the degrees taken includes the following:

- educational six-digit code,
- date when the degree was taken,
- location of the educational institution,
- type of educational institution,
- code of the educational institution.

2.3 Quality of Register-Based Census Statistics

The reliability of register data in Finland was already examined before any decisions were made to introduce a register-based census system. An extensive evaluation survey was also made in connection with the 1990 census. In this evaluation study the same data which were produced register-based were collected questionnaire-based from a sample of 2 per cent of the population. The results are published in a separate publication which is also available in English.

The most important means of monitoring the quality of annual register-based employment statistics is continuous annual quality control using the labour force surveys as reference material. The reliability of the 1995 census data are also assessed in the same way.

The Labour Force Survey serves as reference material for two purposes, *monitoring of the level of the results* produced by the two methods and *monitoring of the extent to which the methods* produce data classified in the same manner.

The Labour Force Survey and the Register-based Employment Statistics characteristically yielded slightly dissimilar results, although no significant discrepancies have been observed in the level of the results or the differences between them from one year to the next. It can be seen from the following figures that although the level of the figures is slightly different, they behave quite uniformly.

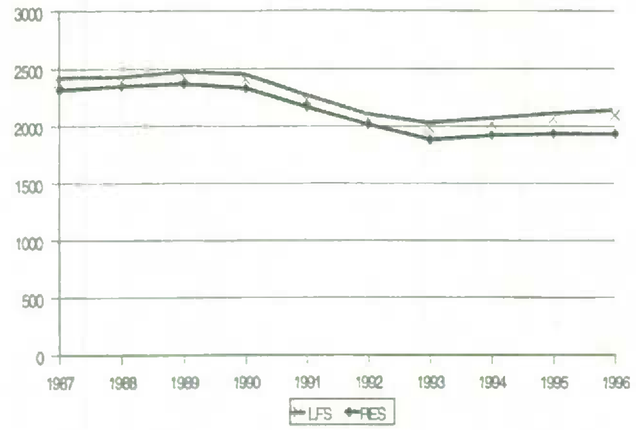


Figure 1. Employed population as indicated in the Labour Force Survey (December sample) and Regional Employment Statistics

The quality of the register-based employment statistics is also monitored every year by comparing the data at the unit level with the data collected in the Labour Force Survey, the latter being used in the following three ways:

1. to identify errors in data processing,
2. to identify situations requiring a change in decision rules, and
3. to check the level of the results.

Quality control at micro level is in practice performed by cross-tabulating the December sample in the Labour Force Survey (LFS) at the unit level with the corresponding sample in the register-based Regional Employment Statistics (RES) (interview-based and register-based data on the same persons in both). The size of the sample is about 12,000 persons, the reference week being the second week of December in the LFS and the last week of December in the RES. This time difference must be taken into consideration when interpreting the results obtained from the comparisons.

The comparisons are performed by cross-tabulating the register-based and interview-based information on the main economic activity and branch of industry of the population making up the sample. Persons classified in the same manner in both the systems are placed on a diagonal line and those classified differently at points off the diagonal. Comparisons have been made since 1987, when deviations from the diagonal were analysed quite accurately. Should the deviation from the diagonal differ from that of the previous year, the case is examined in more detail by consulting the basic material. If the deviation is due to an error in data processing, the error is eliminated, and if it is caused by changes in the register data, for example, or in the associated legislation, the necessary changes may be made in the regulations governing the processing of the data.

Deviations are often attributable to differences between the register system and the interview method so that it is sometimes impossible to state unambiguously which set of results is correct, *i.e.*, the information gained by means of the interviews cannot necessarily be regarded as entirely correct. It is often the case in practice that the interviewer or interviewee has to make a decision in a situation involving conflicting information whereas the register method has a logical rule for resolving the contradiction. The advantage of the register method in such a case is its logical nature, as a machine will always resolve difficult questions in the same manner whereas persons possessing the same information reach different conclusions.

2.4 Statistical Properties of the Register-Based Census Type Statistics

For example the register-based employment statistics system can offer excellent opportunities for examining the processes by which people move from one state in life to another (for instance how young people enter working life) as they enable main economic activity and employment to be examined at various points in time. The employment statistics system possesses the following advantages as far as description is concerned:

1. **The material is based on total count.** This permits precise classifications (*e.g.*, of education), and also analysis in areal terms. The situation can also be examined at the level of individual persons (*e.g.*, in the years y and $y+1$).
2. **The material is compiled annually**, so that flows between years can be monitored.
3. **The material is based on map coordinates**, so that the statistical units can be located geographically and the material can be used to compile statistics for any set of coordinate-based areal units.
4. **The material includes identification codes.** The status of each person in relation to the labour market, for example, at a given point of time can be examined on the basis of data codes (social security numbers). The material can also be linked with other register files as required for the purposes of the analysis.
5. **The material records all forms of activity,** including simultaneous studying and work, and the holding of a number of jobs at the same time.

REFERENCES

Harala, R. (1995). Continuous quality check of register-based employment statistics. *European Workshop on Using Administrative Data in Population and Housing Censuses*, Statistics Finland and Eurostat, Helsinki.

Harala, R. (1995). Evaluation of the results of the register-based population and housing census 1990 in Finland. *Statistical Journal of the United Nations*, ECE 63-72, IOS Press.

Evaluation study of the 1990 Census, Statistics Finland, *SVT, Population Census*, Helsinki 1994.

Laihonen, A. (1996). Making gold from scrap. *IAOS and Statistics Iceland*, 5th Independent Conference Reykjavik, Iceland.

Ruotsalainen, K. (1996). The register-based employment statistics in Denmark, Finland, Norway and Sweden, Statistics Finland and Eurostat, Helsinki.

Vihavainen, H. (1996). Potential use of administrative registers to alleviate the burden on census or survey respondents. Paper prepared for the third Mondory Seminar 25-26 January 1996.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

DEMOGRAPHIC ANALYSES IN SUPPORT OF A ONE NUMBER CENSUS

J. Charlton[1], R. Chappell and I. Diamond

ABSTRACT

The ONS is researching a 'One Number Census' methodology, ideally to adjust all census micro-data for under enumeration, but at least to adjust Census counts to provide best local authority estimates. This talk will present the demographic analysis sub-project, and give some preliminary results. The most important aim is to provide a benchmark against which the survey-adjusted census counts can be validated. The work described here is complementary to ongoing research aimed at improving annual population estimation methods, and will also provide an assessment of the accuracy of current population estimates procedures. The demographic analyses consist of the following: (1) Cohort analyses based on birth and death records, and different sources of data on migration. (2) Analysis of the types of people missing from the 1971, 1981, and 1991 censuses using data from the ONS longitudinal study, with linked data on over 700,000 individuals who were/ were not present in the 1971, 1981, 1991 censuses. (3) Analyses of sex and age-ratios. Sex ratios were crucial to estimating the extent of the census undercount in the 1991 census.

KEY WORDS: Census; Accuracy; Population; Cohort; Estimate; Demography; Validation.

1. INTRODUCTION

The aim of the one number census project is to provide a methodology that enables the ONS to produce census results that have been adjusted for coverage errors and are consistent, ideally down to small area level. The methodology makes use of a large Census Coverage Survey (CCS) and, possibly, administrative records. The adjusted 2001 census figures will be compared with demographic estimates, and if, at a national level, these come within the range of plausible values the adjusted 2001 census data will form the basis of the population estimates for 2001.

Within the One Number Census project the demographic analyses aim to produce:

- A strategy to develop the best possible rolled-forward population estimates, to provide a benchmark comparison for the 2001 census;
- A check on the accuracy of the adjusted census figures for 1971, 1981, and 1991 which formed the basis for subsequent rolled-forward population estimates;
- Estimates of the "margin of error" for national population estimates, taking into account sampling and non-sampling errors in the source data. This will provide the "plausible range" for assessing the adjusted 2001 census figures.

The demographic research will examine sub-national as well as national estimates. However since the availability of data at the two geographic levels is different, different approaches are required. The research has begun with analyses at the national level, the main subject of this paper. Although current research is concerned with England and Wales, the methodology being developed will later be applied in all UK countries.

2. IMPROVING THE METHODOLOGY FOR THE ROLLED-FORWARD POPULATION ESTIMATES

This Section describes the ongoing research programme that aims to investigate:

- sources of error and ways of improving coverage
- the potential of administrative data for improving estimates for particular population age groups
- the method of extinct generations for adjusting the age distribution for the elderly.

2.1 Each year the Office for National Statistics (ONS) produces population estimates for England and Wales by age and sex, at national level, and at sub-national level. The standard ONS methodology is described in detail elsewhere (Population Statistics Division 1978, 1984, 1991; Charlton, Chappell, Diamond *et al.* 1997). These estimates are produced by rolling forward from the most recent census after allowing for under-enumeration, taking into account the births, deaths, and net migration which have occurred in the intervening period. Thus in year t the population P_t is given by: $P_t = P_0 + \sum_i (B_i - D_i + I_i - E_i)$, where P_0 is the base population and B, D, I and E are respectively the Births, Deaths, Immigrants and Emigrants in each subsequent year. For sub-national estimates there are a few other adjustments for definitional differences, *e.g.*, location of students. When there is a new census, it becomes possible to assess the accuracy of the previous ten years of inter-censal population estimates which could lead to "re-basing the estimates" (Armitage and Bowman 1995). However, re-basing population estimates on the census can only occur if the census figures are accepted as definitive. For reasons explained elsewhere (Diamond *et al.* 1997) this was not the

[1] John Charlton, Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ, United Kingdom; e-mail: John.charlton@ons.gov.uk.

case in 1991. There, under-enumeration was not identified by the census validation survey, but instead had to be estimated by demographic analyses which were based on the 1981 census. Thus the base for the current series of annual mid-year population estimates is the 1981 census.

2.2 Potential errors in the annual mid-year population estimates could arise from inaccuracies in: (i) the base census figures; (ii) birth/death counts; (iii) estimates of international and (iv) internal migration, including data on the armed forces and their dependents. The natural change component of the estimates (births minus deaths) is widely accepted to be reliable – compulsory registration systems have been in force in the country since 1839 and improvements have been introduced over time (Charlton and Murphy 1997). The main source of the international migration data is the International Passenger Survey (IPS) which has as its main aims the collection of data on movements to and from the UK by all travelers. Since migrants are a relatively small proportion of all travelers the migrant sample size is small, about 1,300 immigrants and 600 emigrants in 1994. Thus the sampling error is high with confidence intervals of around ± 10% for each of the in- and out- migration flows for a single year. These errors, and non-sampling errors, are discussed in Section 3. In addition, the Home Office supplies data on asylum seekers and visitor switchers, since these are not available from the IPS. A visitor switcher is a person entering the UK who is admitted as a short-term visitor and then stays for a year or longer. Included in these numbers are students and some asylum seekers. Armed forces personnel and students are mobile populations that also need special care. Methods of improving the quality of these data are being explored.

2.3 Each decennial Census is liable to general under-enumeration and miscounting of particular population sub-groups such as armed forces personnel, elderly people and babies. Birth registration data have traditionally been used to produce population estimates in census year rather than relying on the Census. The elderly also tend to be under-counted in the census and the potential use of administrative data such as DSS pension data is being researched. Data from child benefit statistics may be used as a check on counts of children.

2.4 Internal migration is estimated using administrative registers – the National Health Service Central Register (NHSCR) and electoral roll data. The NHSCR is a system administered by ONS which contains records of all persons registered with a General Practitioner and includes data for National Health Service number, name, date of birth, sex and health authority area. When a person moves and changes their GP there is a re-registration that will be recorded in the NHSCR, but only when a person crosses a health service authority boundary. Changes in the electoral roll within local authority areas are currently used to indicate movements at a lower geographical level. We are evaluating Family Health Service Authorities (FHSA) registers as a potential source of migration data, which will be used if they provide better migration statistics.

2.5 The Government Actuary's Department (GAD) was consulted about using the method of extinct generations (Thatcher 1992; Das Gupta 1990, 1991), and especially estimation techniques for "almost extinct" generations. These techniques are ways of improving the accuracy of age distributions for the elderly. Analyses have shown that discrepancies with ONS population estimates only arise above age 90, where figures are not published at present. However if interest were to prompt publication at higher ages, then these methods may be appropriate for ONS.

Adjustments made to produce a new base for rolled forward population estimates

2.6 The coverage checks following the Census in 1981 were the most thorough that had been made up to that time. These were:

- a Census Validation Survey (CVS) from which it was estimated that the Census had missed 214,000 persons.
- demographic analyses of "aged on" 1971 Census population less deaths and allowing for net migration, together with an estimate of the number of children under ten from birth records, taking mortality and net migration into account. Other administrative records (school roll statistics, child benefit statistics, and statistics on pensioners) were also used. These analyses indicated that 36,000 children had been missed by the Census, in contrast to the post enumeration survey estimate of 10,000. The conclusion was that the adjustment to the Census for under-enumeration to form a base for the mid-year population estimates from 1981 onwards was 240,000 persons missed (214,000 -10,000 +36,000). This is equivalent to an undercount of about 0.4%. However the CVS following the 1991 census failed to detect the full extent of the under-enumeration. Comparisons of 1991 census results with the population estimates rolled forward from 1981 and other demographic analyses suggested a net undercount of about 1.2 million residents (Heady 1994). In the final analysis the population estimates rolled forward from 1981 were deemed to be more reliable at the national level than the Census (Heady *et al.* 1994), and are thus also the basis of estimates from 1991 onwards. Since, potentially, the 1981 CVS, which used a similar methodology, could have suffered from a similar problem, there is a need to confirm that the 1981 Census provides the best base for rolled-forward national population estimates (see later).

3. CORROBORATING THE NATIONAL POPULATION ESTIMATES USING ALTERNATIVE METHODOLOGIES

These analyses were undertaken primarily to establish the best base year for rolled-forward population estimates that will be used as a benchmark for the adjusted 2001 census results, and included:

- Cohort analyses as an independent check, tracing births from 1911 onwards, making allowances for mortality and migration based on "best" estimates of international migration.

- Comparison of cohort estimates with 1971, 1981 and 1991 censuses and population estimates.
- Analysis of people missing in censuses, using the ONS Longitudinal Study data.

Cohort Analyses

3.1 Cohort analysis is a way of producing population estimates which are independent of all censuses. Starting with births in each year following an initial base year, these birth cohorts are then "aged on", subtracting deaths and allowing for net migration, to give an estimate of the numbers who remain. Thus if in year *j* the number of births is B_j then the population from this cohort that remains at the beginning of year *i* is given by: $P_{ij} = B_j + \sum_{k=j...i-1} (I_{kj} - E_{kj} - D_{kj})$, where D_{ij} deaths occur in this birth cohort up to year *i*, and I_{ij} and E_{ij} are respectively immigrants and emigrants born in year *i*. This approach is similar to that used in the current rolled forward estimates, except that the most recent census is not used as the starting point or base. A much longer series of births, deaths and migration estimates is of course required. We have used this approach as a method of corroborating the bases for national rolled-forward population estimates.

3.2 A database was compiled of the available data, comprising:

- birth data for each year from 1911
- single year of age cohort life tables for each year from 1911 compiled by the Government Actuary's Department.
- deaths by single year of age, from 1950
- individual birth and death records, from 1968.

A comparison of cohort results using data where both life tables and actual numbers of deaths (or actual death records) were available indicated that little accuracy was lost through using life tables, and so for consistency these were used for the entire period 1911-1991. Those born in 1911 will be 90 by 2001, so there is little to be gained in using data prior to 1911. The methodology used for reconstructing migration series back to 1911, by single year of age and sex, and assumptions made, are described elsewhere (Charlton *et al.* 1997).

Comparing cohort estimates with 1971, 1981 and 1991 censuses and national population estimates

3.3 Using these data, comparisons have been made between:

- cohort analysis results, without taking into account the effect of migration;
- cohort analysis results adjusted for migration effects;
- unadjusted census data for 1971, 1981, and 1991;
- official population estimates.

3.4 Figures 1 to 4 show, for 1991, and 1981 the comparison of the cohort approach with census data and population estimates, with and without allowance for the effects of migration. Figures 1 and 2 compare the estimates for 1991, and also show 95% confidence intervals based on IPS sampling errors (see Section 3). Points to note are the difference between the Census counts (data not adjusted for under-enumeration) and the Population estimates (the revised final estimates for 1991). This equals (by definition) the under-enumeration that was assessed for 1991 when the population estimates were made. Also, in Figure 1 a Census undercount of males in their early twenties is clearly visible when compared with the cohort estimates. The cohort estimates are higher for this age group than the official estimates, but this was not the case in 1971 or 1981. Figures 3 and 4 compare the estimates for 1981. Once again it is clear that population estimates produced by both methods are similar. Similar results were obtained for 1971. The difference between census results and both types of population estimates is much smaller than in 1991. This gives an indication of the likely greater accuracy of the 1981 census. There were more people found in the census than the cohort analyses predicted for ages 30-49 in 1971, 40-59 in 1981, and 50-69 in 1991. This could be because the migration figures failed to include a large number of immigrants in the 1960s, when there was a large net inflow (Coleman and Salt 1992). This will be investigated further using the Labour Force Survey.

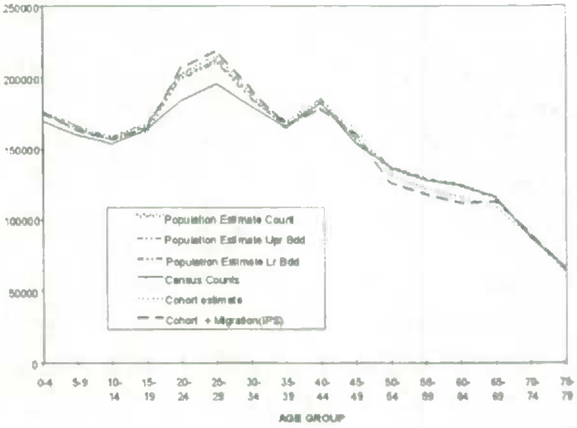


Figure 1. Comparison of rolled-forward population estimates with cohort population estimates (Males - 1991): Migration - IPS

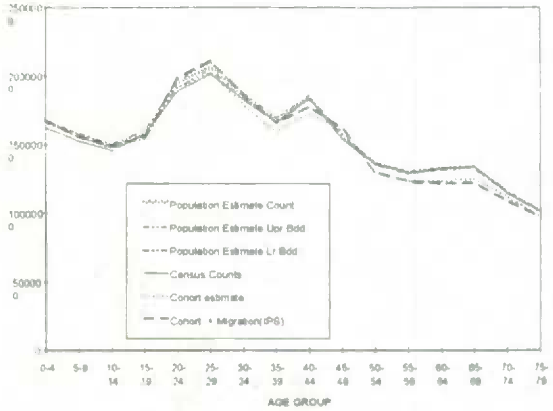


Figure 2. Comparison of rolled-forward population estimates with cohort population estimates (Females - 1991): Migration - IPS

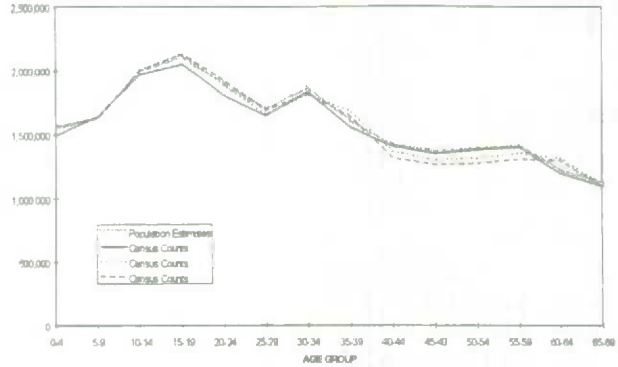


Figure 3. Comparison of rolled-forward population estimates with cohort population estimates (Males - 1981): Migration - IPS

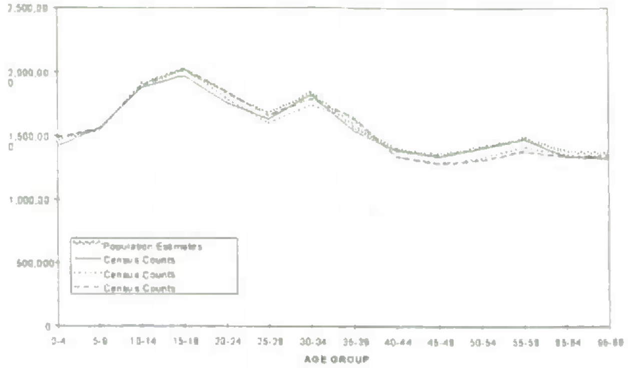


Figure 4. Comparison of rolled-forward population estimates with cohort population (Females - 1981): Migration - IPS estimates

3.5 A cohort approach is independent of censuses whereas the official population estimates are essentially based on a census. However the cohort approach requires migration estimates spanning a greater number of years. Results from the cohort analyses provide population estimates closely in line with 1971 and 1981 Census data, and most similar to the official population estimates for 1981. This corroborates the official estimates at national level, and suggests that 1981 official population estimates provide the best base from which to roll forward population estimates for the 2001 census benchmark.

An alternative check on which census provides the best basis for rolled-forward population estimates – Analysis of census non-respondents in the Longitudinal Study.

3.6 The ONS Longitudinal Study (LS) consists of linked census and vital registration data on a one per cent sample of the resident population of England and Wales. Selection into the LS is by birth date, and the study was designed as a continuous, multi-cohort study, with subsequent samples being drawn at each census, using the same selection criteria, and linked into the dataset (Hattersley and Creeser 1995). It includes the 1971, 1981 and 1991 Censuses, and there are no adjustments for under-enumeration. Each year, new members are entered by virtue of being born on LS dates or by immigration (if born on LS dates) and exited by death or emigration. LS members are also traced to the National Heath Service Central Register (NHSCR). The data on an LS member include everything collected in the censuses.

3.7 The LS can be used to identify the types of people who are apparently missing from censuses, analogous to the Reverse Record Check used by Statistics Canada (Burgess 1988). There are some caveats, however:

- those missing may merely be not linked, *e.g.*, because they have the wrong date of birth on a census form. Birth dates are more likely to be correct on the NHSCR than on the Census. The census form is usually filled in for the household by one individual, who may record the date incorrectly;
- those missing may have been temporarily out of the country as part of an absent household on Census night;
- the LS does not cover Scotland, so those who move there will not be included in the linkage;
- the NHSCR records are only amended when people re-register with a different GP or de-register, for example when leaving England and Wales. Young healthy men may not always re-register promptly with their GP after moving, and those leaving the country will probably not de-register with the NHS.
- When an individual is absent from a census no data are available at that point in time. Information on age, sex, and country of birth will be available from other censuses, however.

3.8 Whilst bearing these caveats in mind, Figures 5 and 6 show estimates of those missing in 1971, 1981, or 1991 but present in the other censuses. These data exclude those who died or were known to have immigrated. Only individuals traced at NHSCR in 1981 and 1991 were included. LS "undercounts" are higher than the undercounts achieved in the censuses, partly because "missing" could also be due to non-linkage of records. Of the three censuses, 1981 had the smallest proportions of missing individuals. Also, young men (and to a lesser extent young women) aged 20-30 were most prone to be missing in the LS. This was the group suspected of being most prone to under-enumeration in the 1991 census. The results also give a feel for the relative magnitude of the under-enumeration problem in 1991.

3.9 We have considered whether it would be possible to use the LS to help to estimate the types of people missed by the Census in 2001. However a verified LS database is only likely to be available one and a half years after 2001. Our provisional view is that any potential benefits of using the LS would be unlikely to warrant making a case for a speedier link.

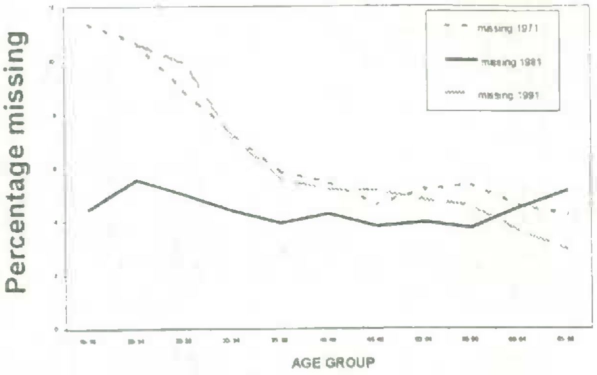


Figure 5. Percentage of women missing in 1971, 1981 or 1991 but present in the other two censuses.
Source: Longitudinal Study, England and Wales

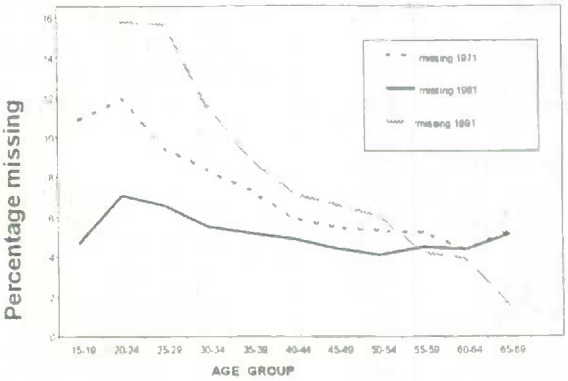


Figure 6. Percentage of persons missing in 1971, 1981 or 1991 but present in the other two censuses Males.
Source: Longitudinal Study, England and Wales

4. ESTIMATING THE "MARGIN OF ERROR" FOR THE ROLLED-FORWARD POPULATION ESTIMATES

The main source of error in population estimates is the data on migration. This Section covers:

- Estimation of sampling errors
- Estimation of non-sampling errors

4.1 As stated in 1.2 above, the key source of error in rolled-forward population estimates is the data on migration. International migration data are derived from the IPS, supplemented by administrative data on asylum seekers/ visitor switchers, and military personnel and their families. The non-migration sources of error in national population estimates can be considered *de minimus*. The rest of this Section is concerned with migration data from the IPS for the period 1975 to 1994, which was readily available for analysis.

4.2 The IPS produces estimates of the number of migrants to and from England and Wales each year, based on stated intentions at the time of interview. It does not cover migration with the Irish Republic, and includes very few asylum seekers or visitor switchers. Asylum seekers usually go through different entry channels and are missed by the survey. A "visitor switcher" refers to a person entering the UK who is admitted as a short-term visitor and then stays for a year or longer, including students, and some asylum seekers. The sample is small (currently under 2,000 immigrants and emigrants are interviewed each year). Figure 7 shows the estimates with 95 per cent confidence intervals for males, for all ages combined (the graph for females is similar). The standard errors have been calculated based on the sampling fractions, and allowing for a survey design effect of 1.2. The sampling error increased from the early 1960s when the sampling fraction was reduced.

4.3 If the benchmark population estimates for the 2001 census are rolled forward from 1981, then 20 years worth of IPS data will have been used in the estimates for 2001, so the confidence intervals for migration by single year of age will be similar to those in Figure 8. From the point of view of making the estimates there is no difference between producing incremental estimates each year and using 20 years of migration data in one go to get from 1981 to 2001. However "final" as opposed to "provisional" versions of data could be used, which should increase accuracy. The confidence intervals may be reduced in relative terms by pooling the samples. The effect on the population estimates of applying the confidence intervals obtained from the IPS for 1981 to 1990 is shown in Figures 1 and 2.

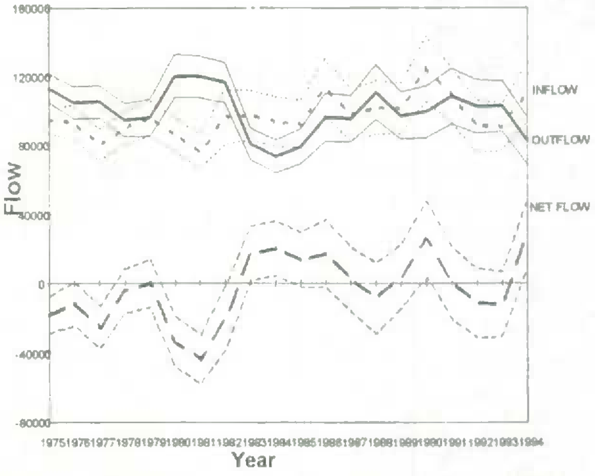


Figure 7. Total Male Migration 1975-94 (Source: IPS Survey) England & Wales Estimates and 95% Confidence Intervals

Estimation of non-sampling errors for the IPS

4.4 As with any survey, in addition to sampling errors, the IPS also has non-sampling errors. People interviewed at ports of entry are asked about their intentions. They may not necessarily tell the truth, or may change their minds as to how long they will stay. It is suspected that the IPS picks up a higher proportion of emigrants than immigrants because of language problems, but it would be an expensive study that could quantify such a tendency. Also more foreign tourists appear to arrive than leave, perhaps due to the greater length of the exit questionnaire. More British tourists appear to arrive than leave. In the IPS a correction factor is applied to compensate for this for tourists but is not used for migrants. The IPS only monitors arrivals and departures during the daytime, and the assumption is made that those on night flights are similar. Some sources of bias in the IPS are reviewed by Bulusu (1991).

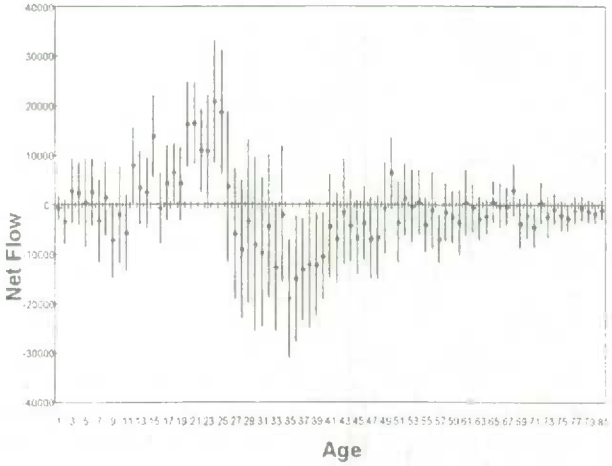


Figure 8. Net Male Migration by age 1975-94 (Source: IPS Survey) England & Wales Estimates and 95% Confidence Intervals

4.5 In order to obtain some estimate of the possible size of non-sampling errors in the IPS the survey data are being compared as far as is possible with alternative estimates of migration, based mainly on:

- The Labour Force Survey (LS), an annual survey which collects country of birth data as well as the date of arrival of the individual in the UK.
- The Sample of Anonymised Records (SARs), a 2% sample of the 1991 Census, used to estimate the immigrant and resident native populations in England and Wales as at 1991.
- Longitudinal Survey (LS), a study containing linked census and vital events data on a 1 per cent sample of he population of England and Wales, which started with a sample drawn from the resident population of England and Wales enumerated at the 1971 Census. The LS contains country of birth data that can be used to estimate the immigrant and resident native populations in 1981 and 1971, as with the SARs.

4.6 These sources of data are being used to produce estimates of migration to check the estimates produced from the IPS data. Two types of comparison are required – net immigration up to a census date for comparison with the IPS plus historic time series data used with the cohort analyses, and short-period immigration for checking the 10- or 20-year immigration data that are used with rolled- forward estimates. The research is in progress.

5. SUB-NATIONAL ANALYSES

Sub-national analyses include:

- Identifying potential improvements in sub-national migration data (*e.g.*, through using FHSA data)
- Identifying locations of transient population groups, *e.g.*, armed forces and students.
- Identifying hard-to enumerate areas, and methodologies for detecting undercounts in certain types of area (sex-ratio analyses *etc.*).

5.1 Methodologies for detecting undercounts in different types of areas

In 1991, sex ratios were examined for ten groups of local authorities (Inner London, Outer London, major metropolitan areas, other metropolitan areas, cities, remote areas, mixed urban rural, new towns, resort/retirement, and one other). The strategy used in 1991 was, for each group, to estimate target sex ratios for 1991 based on their sex ratios in 1971 and 1981 together with national change between 1981 and 1991. Using these target sex ratios it was then possible to estimate age sex specific under-enumeration for each group. However, as the national estimates in 1991 were based on target sex ratios it is somewhat implausible to use the same target sex ratios in 2001 as there is likely to have been real change over the twenty years from 1981. Current research is focussing on the potential to use sub-national sex ratios as a check on the accuracy of census counts for different types of area and to explore the extent to which such sex ratios change over time. Data which adjust for census undercounts for small areas in 1991 (Estimating with Confidence, Simpson *et al.* 1995) could be used as one benchmark of what the sex ratios for particular types of areas should be (although based ultimately on 1971 and 1981 ratios). One approach that may be explored is to use cluster analysis to classify different types of areas according to their age/sex distributions and other characteristics. Then if they do not fit the expected pattern in 2001 this may signal a potential problem. This work is only at an initial stage. It may, in addition, be used to help inform the choice of stratifying variables for the census coverage survey.

6. CONCLUSIONS

The cohort analyses confirm that the adjusted 1981 census forms an appropriate base population for estimates up until the 2001 Census. For the national level confidence intervals for the 2001 population estimates can be calculated based on the sampling error inherent in the International Passenger Survey, and for the total population these amount to about ± 50,000. Further work is required to quantify the potential extent of non-sampling error, and this is being estimated by comparing IPS data with other sources for migration. Quantified sampling and non-sampling error will enable plausible national population ranges to be constructed, against which the 2001 adjusted census results can be validated. The extent to which further demographic analyses are required at a sub-national level is currently under consideration.

REFERENCES

Armitage, B., and Bowman, J. (1995). Accuracy of rolled forward population estimates in England and Wales 1981-1991. *OPCS Occasional Paper,* 44.

Bruce, S., and Elliot, D. (1993). *Sampling Errors and Sample Optimisation on the IPS Vol. 1: Methods, Vol 2: Results* Internal ONS Reports.

Bulusu, L. (1991). A review of migration data. *OPCS Occasional Paper*, 39, London, OPCS.

Burgess, R. (1988). Evaluation of reverse record check estimates of undercoverage in the Canadian census of population. *Survey Methodology*, 14, 137-156.

Charlton, J., and Murphy, M. (1997). *The Health of Adult Britain 1841-1994*, 1, 2. ONS/TSO.

Charlton, J., Chappell, R., Diamond, I., Spencer, C., and Colman, S. (1997). Demographic analyses for a one number census. Paper ONS (ONC(SC))97/04 for One Number Census Steering Committee of June 1997.

Coleman, D.A. (1987). United Kingdom Statistics on immigration: development and limitations. *International Migration Review*, 21, 1139-69.

Coleman, D., and Salt, J. (1992). *The British Population*. Oxford University Press, Oxford.

Das Gupta, P. (1990). Reconstruction of the age distribution of the extreme aged in the 1980 census by the method of extinct generation. Paper presented at the 1990 Joint Statistical Meetings, Anaheim, California, August 6-9 1990.

Das Gupta, P. (1991). Reconstruction of the age distribution of the extreme aged in the 1980 census by the method of extinct generation. *1990 Proceedings of the Social Statistics Section, American Statistical Association*, 154-159.

Diamond, I., Teague, A., Thorogood, D., Brown, J., Buckner, L., Codd, W., Chappell, R., and Charlton, J. (1997). Developing a one number census in the United Kingdom. To appear in Proceedings: Symposium 97, New directions in Survey and Censuses, Statistics Canada, November 1997.

Hattersley, L., and Creeser, R. (1995). Longitudinal Study 1971-1991. History, organisation and quality of data. LS no. 7. OPCS/ HMSO, London.

Heady, P., Smith, S., and Avery, V. (1994). *1991 Census Validation Survey: coverage report. OPCS.*

International Migration 1994 Series MN No. 21 ONS.

Mitchell, B.R. (1992). International historical statistics. *Europe 1750-1988.* Third Edition, McMillan Press, London.

OPCS 1981 Census General Report. *HMSO.*

Population Statistics Division (1978). Population estimates 1971-77. *Population Trends,* 13, 10-12, 1978.

Population Statistics Division (1984). Population estimates 1961-81. *Population Trends*, 35, 30-33, 1984.

Population Statistics Division (1991). Making a population estimate in England and Wales. *OPCS Occasional Paper 37.*

Population Statistics Division (1993). How complete was the 1991 Census? *Population Trends,* 71, 22 - 25.

Rocke, L., and Goodwin, G. (1996). Developing the IPS Sample Design. *Internal ONS Interim Report.*

Rowntree, J.A. (1990). Population estimates and projections. *Population Trends*, 60, 33-34.

Simpson, S. (1993). Measuring and Coping with local under enumeration in the 1991 Census. Paper to the conference on Research in the 1991 Census, Newcastle. Paper available from the author.

Simpson, S., Tye, R., and Diamond, I. (1995). What was the true population of local areas in mid 1991? *Working Paper10, Estimating with Confidence Project*, University of Southampton.

Simpson, S., and Dorling, D. (1994). Those missing millions: implications for social statistics of non-response to the 1991 Census. *Journal of Social Policy*, 23(4), 543-67

Thatcher, A.R. (1992). Trends in numbers and mortality at high ages in England and Wales, *Population Studies*, 46, 411-416.

Werner, B. (1984). Infants aged under 1 in the Census, 1861-1981. *Population Trends*, 38, 18-24.

SESSION I-3

New Directions in Data Collection

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

THE NATIONAL LONGITUDINAL SURVEY OF CHILDREN AND YOUTH LOCATOR TEST

S. Michaud, Y. Clermont, Y. Morin and M.-N. Parent[1]

ABSTRACT

The National Longitudinal Survey of Children and Youth is a longitudinal survey gathering information on children and their environment. In 1994/95, a sample of children aged between 0 to 11 years of age was selected for the first collection. The respondent sample is to be followed biennially. Information is collected mainly from a home interview. There is however information collected from the school teachers and principals. Analysis of the first collection data showed a ceiling effect for the mathematics computation test administered in the school. A number of steps were done to try to correct for the ceiling effect. One of them was the development of a locator test in the home interview. The locator test was developed to assess broadly the child abilities in order to select the appropriate level of the test to be administered to the child in the NLSCY school follow-up. While locator tests are regularly used in test administration, the development of the locator test for the NLSCY had to meet the specific requirements of a household survey and within considerable interview time constraint. This paper discusses the need for a locator test, its development, its implementation in the second NLSCY data collection and examines preliminary results, since collection of cycle 2 ended in June 1997.

KEY WORDS: Locator test; Achievement test; Longitudinal survey of youth.

1. INTRODUCTION

Under the federal government's "Brighter Futures" initiative, Human Resources Development Canada (HRDC) is responsible for the "What Works for Children – Information Development Program". As part of this program, HRDC and Statistics Canada have developed the National Longitudinal Survey of Children and Youth (NLSCY). The survey's purpose is to develop information for policy analysis and program development on critical factors affecting the development of children in Canada.

The first collection took place in 1994 and 1995 and 22,831 children ranging in age from newborn to 11 years were surveyed. This formed the longitudinal sample of the NLSCY. Plans are to interview these children every two years until they grow to adulthood. The interview information is obtained primarily from the person most knowledgeable (PMK) about the child (usually the mother). The PMK provides information about the parent(s) and children (with a maximum of two children per household). In addition, an interviewer-administered assessment to measure receptive vocabulary (the PPVT) is done for children four and five years of age. Children 10 years and older also fill out a self-completed questionnaire. At the time of interview, parents of children at school are requested the permission to contact the school teacher and principal to get information about the child and the school environment (via a mailed questionnaire). For children in grade 2 and up, an achievement test is also administered in the school. In cycle 1, the achievement test focused on mathematics computation. The test was fairly short (between 10 and 15 questions) to try to minimise response burden. A same version of the test was administered to two grade levels (so the same test was administered in grades 2 and 3, grades 4 and 5 ...). Analysis of the results indicated problems with the test and solutions were examined to correct results in cycle 2. Some of the tools for correction included the development of a locator test. Section 2 will discuss the problems with the cycle 1 collection, section 3 will describe the development of the locator test, section 4 will present preliminary results from the cycle 2 to evaluate if the locator test helped and section 5 will give general conclusions and future directions.

2. CYCLE 1 SCHOOL COLLECTION

In cycle 1, a mathematical computation achievement test was administered as part of the school component of the NLSCY. Students in grades two and over completed a shortened version of the Mathematics Computation Test from the standardised Canadian Achievement Tests, Second Edition (CAT/2). The CAT/2 test (developed by the Canadian Test Center) has a whole set of measures of achievement in basic skills – reading, spelling, language, study skills and mathematics. Each level of the tests is related to specific grade, and the tests are equated to relate the scores of the tests from grade 2 to grade 12. While in general, the full version of the CAT-2 on mathematics computation had between 26 to 40 questions, for response burden reasons, a shorter version of the test, targeted to last 15 minutes, was developed for the NLSCY. Initial design

[1] Sylvie Michaud, Yvan Clermont, Yves Morin and Marie-Noelle Parent, Special Surveys Division, and Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

had planned to administer the same test for two grade levels. However, after the first collection cycle, the analysis of the results revealed two problems; a problem with the response rate and a ceiling effect.

2.1 Response Rate to the School Component

In cycle 1, the response rate in the home component was of 83%. However, a consent from the parent and from the school board was required to administer a test in the school. Operational difficulties have been encountered. And since the test was part of a mail-out survey, it also had a non-response component to it (tests not received, or not useable...). Even if individually, none of the factors had a bad response rate, cumulated, they lead to a response rate of approximately 50% to the school component (of responding children to the home interview). Such a large non-response creates problems for the analysis. For example, growth curves require three measures in time. The large non-response implies that in a lot of instances, a fourth collection will be required to be able to calculate them. Studies were done to assess if imputation was possible for the missing test score. Analysis showed that not enough information was available to do a proper imputation (Willms 1996). In longitudinal surveys, it is often observed that once collaboration has been sought from a respondent, one of the main non-response becomes the inability to trace people who move, and that the number of refusals decrease in time (except if an important aspect of the survey design changes). In the school component, children may have changed school (like from switching from elementary to high school) and teachers change. So, that longitudinal collaboration from teachers is likely not going to be there, and refusals could likely be independent between collection cycles. The increase in response rate, but also procedures for correcting for non-response were though important considerations in the design of cycle 2.

2.2 Ceiling Effect

In addition to the non-response to the test, a ceiling effect was observed, especially for certain grades of the test (see Table 1). The ceiling refers to the number of children who obtained a perfect score in the test. A ceiling effect is observed when too many children have a perfect score. This creates difficulties in properly measuring growth. A number of explanations were used to explain this effect. First tests were short (children in grade 2 and 3 had only 10 questions and the ones for grades 4 and up had 15 questions). Secondly, the same test was administered for two grade years (2 and 3), (4,5), (6,7). The ceiling effect was most apparent for the highest grade within a pair (grade 3, 5, 7). Finally, curriculum is a provincial jurisdiction in Canada and the number of years required to complete high school varies by province (for example, in Quebec, there are six years of elementary school while there are seven years in Ontario). So especially for older grades, to create an appropriate grade level test that is also a short test is quite a challenge.

Table 1
Ceiling effect observed in cycle 1 mathematical test

Grade	% of children with a perfect score										
	Total	Nfld	PEI	NS	NB	Que	Ont	Man	Sask	Alb	BC
2	10.8	6.9	12.5	15.7	17.2	15.5	6.9	6.2	8.6	11.1	12.7
3	38.2	47.2	34.5	45.6	40.3	56.5	28.0	32.6	23.1	45.6	42.6
4	3.3	0.0	0.0	5.7	1.5	7.2	3.6	3.1	1.7	0.0	4.7
5	14.6	8.2	3.8	10.3	17.2	30.0	14.3	7.0	13.3	20.2	7.5
6	4.7	6.7	3.0	6.0	0.0	14.5	3.2	1.9	2.1	5.5	3.6

2.3 Corrective Measures for Cycle 2

A number of actions were taken to correct for the mentioned problems in cycle 2. Intense public relations was done to seek participation with school boards. In addition, collection procedures were improved. All the tests were increased to fifteen questions. And a single test was developed for each grade. However, it was felt that this would not solve the problem of different curriculum by province. To correct for this, a longer version of the test was required. However, this requirement was balanced against the decision to measure a new dimension of achievement in cycle 2; reading comprehension. To solve the problem it was decided to build a locator. The locator test is a fairly short test, administered in the home. The result to the locator test, added with other available information, will determine if in the school, the child should be administered the test for his/her grade or the test of the next grade. The locator could also potentially be used to do imputation of missing values.

3. CREATION OF THE LOCATOR TEST

To maintain consistency with the tests in the school, it was felt that a locator should be based on items coming from the CAT-2. There was also a huge database that was available that allowed some tests for the creation of a locator. However, a number of hypothesis had to be made in the creation of the locator. The CAT-2 has been developed in English only. So data was available in all provinces but in English only. For the creation of the locator the assumption that the relations that were observed would hold in French in the same way as in English, and that observed differences were due to a difference in curriculum. Since the tests in the CAT-2 are achievements tests, items are based on the curriculum. So it was assumed that the ceiling effect that varied between provinces was due to the differences in curriculum. There were also criteria in the design on the locator:

- Items to predict reading comprehension were selected from the CAT-2 items in reading vocabulary. Items to predict mathematics computation were selected among items from the mathematics concepts and applications. To do the analysis, an artificial score was created, based only on the subset of questions that would been administered in the test sent in school [CTC 1989].

- A very easy item always had to be present as the first item, not to discourage children.
- Items had to cover a range of item difficulty and discriminating levels.
- Items had to be "translatable" (this excluded homonyms from the vocabulary).
- The minimum level of difficulty had to be reasonable (there was no official threshold but items with an item difficulty that was too low *i.e.* less then .40 were in general excluded because the locator is administered between December and February while normally the difficulty level are based on a test administered in April.
- Finally, for the vocabulary questions, if certain items were grouped, the whole group of items had to be selected in the locator.

A locator test had to be created for each grade level that could be applicable to a child in the sample. That meant that seven locator tests were generated, one for each grade from grade 2 to grade 9. The children would be administered the locator for their appropriate grade level (age was used as a proxy when grade was unknown). To measure the two dimensions of the tests in school, the locator had two sections, one on vocabulary and one on mathematics concepts. An example was provided at the beginning of each section.

Normally, we would have like to administer the locator to a group of students, and then to study the properties. However, because of temporal and budget constraints, studies were done only on a file of test results available from CTC (with test results for a sample of approximately 100,000 children). The whole CAT-2 had been administered by CTC to 92 school boards. Again assumptions had to be made;

(i) The 92 school boards had to be assumed to be a representative sample of the school boards in Canada

(ii) There were not enough test data in the provinces of Quebec and of PEI to do inferences. Assumptions had to be made that the properties of the items would hold for these two provinces in the same way as in the other provinces.

(iii) Assumptions had to be made that as long as the items are kept intact (*i.e.*, a question is not modified), the items would have the same property, whether they are in the CAT-2 or in a shorter test.

(iv) It was assumed that the conditions at home to administer the locator would not impact significantly on the results. This assumption is very hard to validate and this may likely not hold. Interviewers will also be given suggestions on how to administer the test at home, and the conditions of the administration of the test are recorded.

(v) Because the reading comprehension test was not finalised, four passages were taken as an approximation for the actual test that would later on be created.

The studies were done both at the Canada level and at the provincial level. To do some validity checks, the CTC file was divided randomly in two sub-groups. Two tests were initially created. In the first test, items that deemed appropriate in a household and interview and with varying levels of difficulty were chosen. A second test was constructed by picking the items that best predicted the mathematics computation and reading comprehension scores. The number of questions was restricted by the actual interview time that could be allowed in the household interview to do this locator test. Reliability were compared for different number of questions. Finally, a third test was created by combining the first two tests. The three tests were compared. A summary of the comparisons of the three models is given in Table 2.

Table 2
Tests for the different designs of the locator; for grade 4

Model	Reliability			mean difficulty (10 items test)
	10 items	12 items	8 items	
operational	0.73	0.76	0.68	5.9
regression	0.70	0.73	0.65	5.3
mixed	0.77	0.80	0.73	6.0

In general, there were not large differences between the three locators that were compared. The reliability of the test varied between . 70 and .77. The reliability is however too low to measure a sub-component (mathematics concepts or vocabulary). However, since the purpose of the test was mainly for imputation and for determining a grade level, this was judged satisfactory. On average, decreasing the test to 8 questions decreased the reliability by 5%, while putting the test to 12 questions usually increased it by approximately 3%. The locator with the highest reliability for the 10 item test was selected. In half of the grades, the locator that was selected was the one that had been designed with operational considerations, while in the other half of the times, the mixed test was chosen. The model coming from the regression only was never selected because it usually divided items that should have been kept together.

4. ADMINISTRATION OF THE LOCATOR IN CYCLE 2 AND PRELIMINARY RESULTS

In cycle 2, the home interview got a response of 92% for the children in the longitudinal sample. A test in the school was received for approximately 70% of the responding children. The locator was administered at home in cycle 2, and it got a response rate of 92%, which is a reasonable response rate. On average, children took 6 minutes and a half to complete the locator and half of them were able to complete it in less than 3 minutes. Some questions also recorded information on the environment in which the test was administered (were there potential distraction for the child, interruptions…). Rules were then required to decide when the results from the locator should imply that the test of the next grade should be administered at school.

Rules for the locator were not applied to children in grade 2. The curriculum was found to be too different from grade 2 to grade 3 (6 items out of 15 were not in the curriculum) and not enough information was available related to the equating of the test for the different grades to risk to apply the procedure. For the following grades, a distribution of the scores of the locator were examined. At the end, it was decided that children with a score of 9 or 10 out of 10 would receive the next grade level test in school. The reading comprehension test has however been a new test developed for Statistics Canada's cycle 2 collection. Norms do not exist for the test. To be able to equate the test, it was decided that the rules for the locator would be applied to only 75% f the children. Otherwise, if the locator was a perfect predictor instrument, by applying the rule to all the "good" children, none of them would be available in the sample to norm the test in an appropriate grade level and the norming sample would be biased. Since a floor effect had not been observed (the floor effect is observed if too many children have no correct answers), the locator was not used to put children in a lower grade level test.

4.1 Ceiling Effect

Table 3 compares the ceiling effect for the mathematics computation test in cycle 1 and cycle 2. The ceiling effect has been solved in all grades except grade 2 (where a locator was not used). Even if the grade 2 test was lengthened from 10 to 15 question between cycles, the added questions did not completely solved the ceiling effect (in certain provinces in particular).

Table 3
Ceiling effects for cycle 1 and 2

	grade 2	grade 3	grade 4	grade 5	grade 6	grade 7	grade 8
cycle 1	10.8%	38.2%	3.3%	14.6%	4.7%	–	–
cycle 2	9.2%	4.8%	2.6%	1.9%	3.0%	0.4%	1.5%

The rules for the locator allowed the possibility to put children in the next grade level. An assessment was done to see if these children that were administered the next level up had reached a ceiling effect again (this means that for them they should have received a test two grades level higher instead of just one) and also if a floor effect had been obtained for them.

Table 4
Ceiling effects and floor effects for children that received the test for the next level up in the school test

	grade 2	grade 3	grade 4	grade 5	grade 6	grade 7	grade 8
ceiling	0%	0%	0%	0%	1%	0%	0.8%
floor	–	–	4.4%	1.4%	0%	1%	0.8%

This seems to suggest that to send children to the next level up is an appropriate rule and there are no more ceiling effects. There is a slight floor effect for children in grade 4. Further analyses are required to see if the problem can be solved.

Comparisons were also done to see how the children who had received the next level up of the test compared with children of that grade level. The comparisons focused on the average number of correct values, for both the mathematics computation and the reading comprehension test. Results are presented in Table 5. The results suggest that the children who received the next grade level are doing at least as good if not better on average than the children from that level. Except for grades 4 and 5 in mathematics computation, the average score of the children that received the next grade level is always equal or higher than the average of the children from that grade. This suggests that especially for higher grades, the locator identifies children with higher achievement levels.

Table 5
Comparison of the average number of correct answers for children who received the test from their grade level vs. children who received the test from the next grade level.

		grade 4	grade 5	grade 6	grade 7	grade 8
math	grade	8.7	8.7	8.9	6.7	7.9
	grade + 1	7.1	7.0	9.4	8.5	8.5
reading	grade	11.3	10.0	10.4	10.5	10.6
	grade + 1	11.6	11.2	11.6	11.7	10.6

4.2 Non-Response and Imputation

As mentioned earlier, a response was obtained for 92% of the children. Table 6 presents the response pattern for children eligible to receive a test in cycle 1 and cycle 2. Because the mathematics test is a longitudinal measure, to strengthen the analysis, when one of the test score is missing, it could be decided that a score will be imputed. Imputation could also be done at least to analyse the non-response and evaluate its impact on analysis.

Table 6
Response distribution for children in grade 4 and up in cycle 2

Locator	Math cycle 1	Math cycle 2	possible action	%
yes	yes	yes		35.3
no	yes	yes		2.2
yes	yes	no	impute cycle 2	8.1
no	yes	no	impute cycle 2	4.9
yes	no	yes	impute cycle 1	27.3
no	no	yes	impute cycle 1	2.1
yes	no	no		9.1
no	no	no		10.9

The first two rows show that a measure with the two points in time would be available for 37.5% of the children. Twenty percent of children have both of their mathematics score missing (for 9.1%, a locator score could help in a non-response analysis). More than 45% of the children had a score missing in one of the cycles.

Preliminary analysis were done to see how useful the locator could be in an imputation mechanism. The strategy assumed that the imputation of a mathematics computation score could be done using a linear regression model (the imputation was limited here to mathematics computation because reading comprehension was not measured in cycle 1). The analysis were done on the subset of the population that had a locator and a mathematics test to the cycle 2 (for grade 4 and up, they also had to have a mathematics test for the cycle 1). The model was applied separately for each grade level. Each model used variables that had been suggested by experts to be related to a mathematics score. Variables in the model included variables such as mother's education, income, how well the child did at school (parents assessment), province. These variables were labelled as basic information.

Four models were compared. Model 1 restricted to basic variables. Model 2 added the locator score to the basic variables. Model 3 included the cycle 1 mathematics test result with the basic variables. Finally model 4 included the basic variables and both the locator test results and the mathematics score of the cycle 1. The analysis compared the different R2 obtained by the different models. It also compared the predicted "imputed" values with the cycle 2 mathematics score. The R2 values are presented in Table 9 for the four models. C1 refers to cycle 1. The models are done by grade, using the grade in cycle 2 (this means that for grade 2 and grade 3, models 3 and models 4 were not an option since there could not have received a test in cycle 1 because of their age).

Table 9
Comparisons of the four models, for the imputation of a mathematics computation score potentially in cycle 2

	Model 1 (basic)	Model 2 (basic, locator)	Model 3 (basic, C1)	Model 4 (basic, locator C1)
grade 2	21%	25%	–	–
grade 3	16%	17%	–	–
grade 4	29%	32%	45%	45%
grade 5	29%	32%	30%	34%
grade 6	29%	35%	43%	47%
grade 7	28%	34%	33%	38%
grade 8	35%	42%	44%	48%

A number of interesting findings can be derived. First for grade 3, the locator does not add much as a predictive variable to the mathematics computation score. But for the lower grades, the observed prediction levels are much smaller than for higher grades. For grade 4, the single best predictor of the mathematics computation result in cycle 2 is the result in cycle 1. The locator does not add much, if the mathematics computation score is already present. It does help a little when the score from cycle 1 is absent. For grades 5 and up however, the locator seems to bring additional information to the model. As can be seen from the comparison of model 3 and model 4, even if the mathematics score from cycle 1 is available, the information from the locator is also statistically significant and it improves the R2 value by 4%- 5%. So the locator seems to show prospect as an imputation tool, but more for the older grades.

5. CONCLUSIONS

The use of tests at school in a survey environment such as the NLSCY, with the operational and time constraints, posed a number of challenges. The locator test has been one measure proposed to solve some the problems of ceiling effect due to the short tests. Initial results suggest that it worked reasonably well and achieved some of its objectives.

A lot more research is required. There are grades where the locator did not perform as well. No research has been done yet on its impact on reading comprehension. And the longitudinal proprieties of the tests have not been studied. The fact that the survey is in CAI may lead to the development of a locator that becomes eventually more like an adaptive test (where the length and the flows of the questions depend on the response of the child).

REFERENCES

Canadian Test Centre. (1989). Canadian Achievement test. Second edition, Technical Bulletin.

Castonguay, H., Malo, D., and Michaud, S. (1995). Étude sur l'erreur de réponse pour le test de mathématiques de l'ELNEJ. Statistics Canada, internal report.

Germain, M.F., Michaud, S., and Lynch, J. (1996). Using scales in a large survey about children. Paper presented at ASA meeting in Chicago.

Michaud, S., Clark, K., and Morin, Y. (1996). Methodology used for the creation of the locator test for the NLSCY. Statistics Canada, internal report.

Statistics Canada. (1996). Math and reading skills indicator, NLSCY – cycle 2, survey instrument, copyright CTC.

Statistics Canada. (1995). Overview of the instrument, NLSCY.

Willms, D.J. (1996). Indicators of mathematics achievement in canadian elementary schools. *In Growing Up in Canada, Human and Resources Canada & Statistics Canada*, catalogue no. 89-550-MPE, no. 1.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A COMPARISON OF MAIL AND E-MAIL FOR A SURVEY OF EMPLOYEES IN FEDERAL STATISTICAL AGENCIES

M.P. Couper, J. Blair and T. Triplett[1]

ABSTRACT

This paper reports on the results of a study comparing e-mail and mail for a survey of employees in several government statistical agencies in the U.S. As part of a larger study of organization climate, employees in five agencies were randomly assigned to a mail or e-mail mode of data collection. Similar procedures were used for advance contact and follow up of subjects across modes. This paper describes the procedures used to implement the e-mail survey, and discusses the results of the mode experiment.

KEY WORDS: Electronic mail; Mode of data collection; Organizational surveys.

1. INTRODUCTION

With the proliferation of electronic communications in the last several years, electronic mail (e-mail) is an increasingly attractive alternative to mail for surveys of employees in organizations with high penetration of e-mail technology. The major advantages claimed for e-mail over mail are reduced costs and quick turnaround. However, concerns have been raised about issues such as coverage, nonresponse, and measurement error effects of e-mail data collection.

With this in mind, we embedded a mode experiment in an organizational climate survey of employees within several statistical agencies in the U.S. The mode experiment was designed to evaluate the relative quality of the two methods (mail and e-mail) for surveys of federal employees. In this paper we describe the steps taken to implement the survey, and discuss the results of the mode comparison.

2. DESIGN AND ADMINISTRATION OF THE SURVEY

The mode experiment was embedded in an organizational climate survey conducted on behalf of a consortium of federal statistical agencies in the U.S. The study was designed by graduate students in the Joint Program in Survey Methodology (JPSM) as part of the survey practicum. Data collection was undertaken by the students in conjunction with the Survey Research Center (SRC) at the University of Maryland. The overall objectives of the study were to develop and test an organizational climate survey suitable for implementation within federal statistical agencies. The instrument, and the data collected, would be used to benchmark climate within and between agencies of similar composition and function. Details of the survey are reported elsewhere (see Carlson and Rivers 1997). We focus here on the mode experiment.

The survey instrument was developed through several iterations of testing, including two focus groups, several cognitive interviews and conventional pretests. The final instrument consisted of 81 Likert-type attitude items and 10 background items. Nine agencies participated in the larger climate study. The sample was restricted to all permanent employees at these agencies. This included part-time workers, but excluded temporary employees such as coders and interviewers, as well as contract workers.

One of the problems of doing organizational studies is the relatively large sampling fractions required for subgroup analysis. This leads to potential contamination issues (with some employees in an office getting a questionnaire and others not), as well as concerns about providing all employees with an opportunity to voice their opinions about the organization. Thus, we hoped to do a census of all employees in the participating agencies. We could not afford to do this using traditional methods of data collection, which led to the decision to consider e-mail (considered to be much cheaper than mail at the time of our decision). The cost savings would obviously be greater in the large agencies. At the same time, using both mail and e-mail would allow us to test the efficacy of the alternative method of data collection for a survey such as this.

We were assured that all employees at the nine agencies had access to electronic mail, and we were provided with electronic data files containing employee names, office addresses, telephone numbers and e-mail addresses. Given the logistical issues of launching two surveys in each of nine different agencies, we decided to restrict the experimental mode comparison to the five largest agencies. The remaining agencies were given a choice of a single mode. Table 1 lists the number of employees in each of the five agencies assigned to each mode of administration. The imbalance for Agency B is due to the fact that the mail mode was further split between an anonymous and an identifiable group.

[1] Mick P. Couper, Joint Program in Survey Methodology and University of Michigan; Johnny Blair and Timothy Triplett, Survey Research Center, University of Maryland, P.O. Box 1248, Ann Arbor, MI 48106, U.S.A.

Table 1
Sample Sizes for Mode Comparison

Agency	Mail	E-mail	Total
A	2,699	2,969	5,668
B	790	396	1,186
C	266	265	531
D	216	221	437
E	216	215	431
Overall	4,187	4,066	8,253

3. DATA COLLECTION PROCEDURES

In this section we discuss the acquisition and evaluation of software for conducting the e-mail surveys, and describe the general data collection procedures used for both modes. Because of cost and time constraints we decided against developing our own e-mail survey software and instead examined several commercial products for conducting e-mail surveys. An initial review of technical specifications led to an elimination of all but two products.

We first examined Raosoft's "EZSurvey" (see <http://www.raosoft.com>). Advantages of this product included easy development of instruments, an auto-reply feature that facilitated the return of completed questionnaires, the ability to handle various question types, skip functions, and the use of a graphical user interface (GUI). In addition, the software is available for a fixed price, leading to economies of scale and possible use in other studies. However, EZSurvey creates a DOS or Windows executable file, which means the user's operating system needs to be known in advance. We discovered that the size of the outgoing e-mail file approached 1 Mb per sample person which was unacceptable, both in terms of the volume of Internet traffic (over 4 Gb for outgoing messages alone), and because of likely agency restrictions on the size of incoming files. We were also unable to test the auto-reply feature in-house (using Pegasus Mail); nor did it work for any of the seven pretest subjects (technical contacts at the participating agencies). The vendor's initial solution to this problem was to have respondents change their Windows configuration (.ini) file; this would be done by e-mailing all sample persons an executable file to automatically update their system configuration. This was deemed unacceptable. Raosoft then offered to fix the program and have us do the testing. Given that this was three days prior to the start of data collection (which could not be postponed because of external constraints), we had to seek an alternative solution.

We switched to Decisive Survey, a product of Decisive Technology Corporation (see <http://www.decisive.com>). Decisive is a text-based system, so would work on all operating systems. However, the interface is less appealing than a GUI system, and the system does not accommodate embedded logic such as skips. The product is priced on a sliding scale depending on sample size. In addition, no unique identifiers are attached to messages; however, different surveys could be identified with unique authentication markers. This meant that we had to rely on e-mail addresses to match returns back to the frame, something which proved to be quite difficult in practice. Still, we were able to successfully test Decisive Survey with persons at each agency, so decided to proceed with the mode experiment.

The mail survey materials were printed in booklet form, on 8½ by 11 inch paper. The questionnaire was 12 pages long, including a cover with the JPSM logo and title of the survey. An ID number was placed on the back of each questionnaire. A cover letter signed by the director of JPSM, and a reply-paid envelope were included in the packet. The envelopes were individually addressed, and delivered to each agency in bulk for distribution using the internal mail system. Returns were mailed directly to SRC where they were processed and responses keyed.

Similar strategies were used for the e-mail version. However, whereas the items in the paper version were grouped into 14 sections, the e-mail software required all 94 items and sub-items to be numbered consecutively. The closed-ended questions were answered by placing an X (or any character) inside a set of brackets [ ] alongside the option. Open-ended questions were answered by typing within the brackets. A message from the director of JPSM accompanied each instrument. The e-mail messages were sent from SRC, with the return address being *agency*@cati.umd.edu (so each agency's returns came to a different mail queue).

Both e-mail and mail questionnaires were delivered to sample persons on approximately the same day. The survey mailing was also preceded by an advance letter or message from the head of each agency informing their staff of the upcoming survey and encouraging participation.

Approximately one week after the initial mailing, a reminder postcard or e-mail message was sent to all sample persons. Two weeks after the reminder, a second mailing or e-mail message containing a replacement questionnaire was sent to all nonrespondents. Finally, telephone reminder calls were attempted for all remaining nonrespondents about 6 weeks after the initial mailing. No attempt at refusal conversion was made, but replacement questionnaires were offered, and reasons for nonresponse (when provided) were recorded.

A final source of data on the process came from a set of debriefing calls, conducted among respondents from both mail and e-mail treatments. We solicited respondent reactions to the content of the questionnaire and (in the case of e-mail) to the mode of data collection.

4. RESULTS

We have a variety of data sources to evaluate the mode comparison. These include a tracking database in which all transactions (outgoing and incoming mail and e-mail) were logged, a small debriefing study of respondents, reminder calls to nonrespondents, and the substantive responses to the survey itself. We discuss each of these in turn.

First, we examine the response rates by agency and mode. These are presented in Table 2. For each of the five agencies, e-mail produced a significantly ($p < .01$) lower response rate than mail. This finding is consistent with that of most other tests of e-mail versus mail (Bachman, Elfrink and Vazzana 1996; Mehta and Sivadas 1995; Schuldt and Totten 1994; Sproull 1996; Tse 1996). The largest differences in response rate are found for agencies A, D, and E. There are several possible explanations. For Agency A and E, the e-mail addresses were constructed from lists of employee names (following agency conventions such as *last.middle.first@agency*). All other agencies provided e-mail addresses for their employees.

Table 2
Response Rates by Agency and Mode (in percent)

Agency	Mail	E-mail	Difference
A	68.0	36.7	31.3
B	76.1	62.6	13.5
C	74.4	60.0	14.4
D	75.5	52.9	22.6
E	76.4	54.9	21.5
Overall	70.7	42.6	28.1

Another source of the difference may be technical problems with different e-mail systems at each agency. We discovered that Lotus CC: Mail can be set to automatically convert e-mail messages over a certain size (*e.g.*, 20 Kb) into attachments. Both Agency A and D use CC:Mail. However, so does Agency B, which had the highest e-mail response rate. We received several reports from employees from Agency A and D that they received attachments, and didn't know what to do with them. Subsequent investigation suggests that this does not appear to have been a problem in Agency B, and some users at Agency A received the survey as intended (in the body of the message rather than as an attachment). However, the attachment problem appeared widespread at both Agency A and D. As soon as we learned of this, we sent an additional e-mail message to sample cases in these two agencies with updated instructions on how to deal with attachments. Similar problems were not experienced at Agency C (Novell GroupWise) or Agency E (GroupWise or WPMail).

Further evidence for the technical problems caused by the size of the e-mail survey (23 Kb) can be found in the supplement response rates. All agencies were offered the opportunity to include a set of agency-specific supplement questions; only two agencies (A and D) availed themselves of this opportunity. For the mail survey, the supplements took the form of a single sheet insert printed on color paper. For e-mail, the supplement questions were sent in a separate e-mail message. The main and supplement response rates for these two agencies are presented in Table 3. It can be seen that if response rate was defined as *any* completed questionnaire (main or supplement), the overall response rate for Agency A would increase by 19.5% (to 56.3%), while that for Agency D would increase by 16.3% (to 69.3%), whereas the mail response rates would remain unchanged. These new rates are close to those for e-mail in the two agencies (B and C) which did not experience technical difficulties receiving the e-mail questionnaire.

However, even taking these supplement return rates into account, we still find consistently lower response rates for e-mail relative to mail across all agencies. It is thus important for us to explore what reasons there may be for the response rate differential, and what effect this may have on the quality of the data obtained.

In a *tracking database* of all returns, SRC staff noted which cases required special attention for a variety of reasons. For an e-mail survey to be cost-effective, the goal is to minimize clerical activity required. Table 4 shows the various types of clerical action that were required for those e-mail questionnaires that were returned. These may also be indicators of the types of difficulty experienced by e-mail sample persons. The first column shows the percentage of returns that were received as an attachment to an e-mail message, while the second column denotes messages that required decoding. In both cases there is a variation across agencies, suggesting different technical approaches to sending e-mail. From the third column, we

Table 3
Main and Supplement Response Rates by Agency and Mode (in percent)

	Mail			E-mail		
Agency	Main plus supplement	Main only	Supplement only	Main plus supplement	Main only	Supplement only
A	64.9	3.0	0.0	31.2	5.6	19.5
D	72.7	2.8	0.0	46.2	6.8	16.3

Table 4
Types of Clerical Action Required for E-mail Returns (in percent)

Agency	Attachment	File coded	WP file	Needs Edit	Any clerical action	(n)
A	8.2	14.2	17.7	37.9	57.2	(1,091)
B	1.7	0.4	4.3	13.8	20.9	(239)
C	23.7	1.3	23.7	0.0	23.9	(159)
D	12.0	18.5	21.3	19.7	55.6	(117)
E	20.9	0.9	20.9	0.0	20.3	(118)
Overall	9.5	9.9	16.1	27.2	46.5	(1,724)

can see that about 16 percent of cases overall were completed using a word processor or text editor. Noting that there is overlap in these types of problem (all three could occur on a single return), about 21% of all e-mail returns did not make use of the reply feature. Overall, about 3.9% of the e-mail respondents printed out the questionnaire and mailed it back (included in the above figure). Furthermore, a large number of cases required additional editing before the data could be appended to the database. The most common reasons were the X placed outside of the brackets, or one of the brackets deleted. The fourth column shows that about 27% of cases required such editing, but again there is substantial variation across agencies. The final column identifies the percentage of returned e-mail surveys that required *any* clerical action before appending to the database. The high overall rate suggests a great deal of attention was required for the e-mail cases, potentially nullifying the savings in post-collection processing. In addition, the two agencies with the lowest e-mail response rates also exhibit the highest rates of clerical action among returns, again suggesting that technical difficulties experienced by sample persons could have affected the response rates.

As noted earlier, we also conducted a set of telephone *debriefing interviews* with those who returned their questionnaires. A total of 694 sample cases were selected from among the respondents, using several replicates to include both early and late returns. The sample was evenly split between modes, and Agency A was undersampled because of its relatively large size.

Interviews were conducted by JPSM students, ensuring that no student called a respondent from their own agency or known to them. A small portion of the calls were conducted by members of an undergraduate survey methods class. An overall response rate (complete/eligibles) of 77.2% was obtained. The cooperation rate (complete/contacted) was 90.5% (including callbacks) or 98.3% (excluding callbacks). This yielded a total of 244 mail and 256 e-mail respondents who completed the debriefing. While we caution about generalizing from this group of cooperative respondents to the full sample, we can nonetheless gain some insight into the process of data collection from these interviews.

E-mail respondents were asked what method they used to complete and return the questionnaire. Their responses are shown in Table 5. These findings parallel those shown in Table 4, and suggest that the difficulty of replying to the survey differed across agencies. In Agency A and D, about two-thirds of respondents used a text editor or word processor to complete the survey, whereas the survey was designed to be completed using a reply function within e-mail.

We asked both sets of debriefing respondents (mail and e-mail) to estimate how long they took to complete the survey: e-mail respondents reported taking significantly longer ($p < .01$) than mail respondents (28.3 minutes versus 22.5 minutes). While the difficulties in completing the e-mail survey reported above may have contributed to the increased time, there are no significant differences in the reported time of e-mail completion across agency. In other words, even for those agencies which did not appear to experience technical problems, e-mail was still reported to take longer to complete than mail.

Table 5
Method of Reported E-Mail Return by Agency, Debriefing Respondents (in percent)

Agency	Reply Function	Text Editor	Other	(n)
A	20.5	67.0	12.5	(88)
B	55.1	37.7	7.2	(69)
C	64.3	31.0	4.8	(42)
D	6.9	65.5	20.7	(28)
E	78.6	14.3	7.1	(29)
Overall	41.8	47.3	10.9	(256)

One of our initial concerns about e-mail was related to confidentiality. Respondents were being asked to give their candid views on their employers, and the non-anonymity of e-mail may contribute to a reluctance to complete the survey in this mode. We asked debriefing respondents how easy they thought it would be for (a) their supervisors and (b) anyone else in their agency to get access to their (mail or e-mail) responses. Using a 10-point scale where 1 means very easy to get access and 10 is very difficult (thus a high score means low confidentiality concern), the average responses by mode are presented in Table 6. Neither of these differ significantly ($p > .05$) by mode. Thus, among the debriefing respondents at least, there does not appear to be greater concern about the confidentiality of their e-mail responses.

The *reminder calls* may give us further insight into the reasons for nonreturn of the surveys. Toward the end of the study, we attempted to contact all remaining nonrespondents to encourage participation. However, given the high level of nonresponse, time and funds did not permit a concerted effort to contact every nonrespondent. A one-call rule was implemented to ensure that at least one attempt was made for every case. The outcomes of the reminder call attempts are presented in Table 7. The "other" category includes wrong numbers, sample persons who had left the agency, and so on.

Table 6
Mean Response to Two Questions About Access to Survey Responses by Mode, Debriefing Respondents

Question	Mail	E-mail
Supervisor access	6.15	6.61
Access by others in agency	6.32	6.36
(n)	(244)	(256)

Table 7
Outcome of Reminder Calls by Mode

	Mail		E-mail	
Outcome	Percent	(n)	Percent	(n)
Talked with sample person	46.7	(433)	43.2	(964)
Call back	24.7	(229)	30.6	(683)
Left message	12.6	(117)	17.0	(377)
Other	16.0	(148)	9.2	(207)
Total	100.0	(927)	100.0	(2,231)

In Table 8 we present the results of the call for those persons with whom we made contact. First, e-mail contacts were more likely to say they were not going to the return the questionnaire (37.3% versus 22.9%). Among these, almost half (45.8% of the refusers and 17.1% of all those contacted) claimed that they did not receive the questionnaire by e-mail or had lost or deleted the message, but did not want to be sent another. This appears to be less of a problem with the mail questionnaire, suggesting that delivery of an e-mail instrument may be more problematic than that of a mail instrument. Second, 2.7% of the e-mail contacts (or 7.2% of those who said they would not respond) reported difficulties editing the instrument as a reason for nonreturn; as one would expect, no mail contacts reported this reason. Third, the most interesting finding from this table is that the proportion of contacts mentioning confidentiality as a reason for nonreturn does not differ by mode. In fact, 6.1% of the mail contacts who do not plan to respond mentioned confidentiality concerns, compared to 3.2% of e-mail contacts who planned not to respond. These findings again suggest that technical difficulties, rather than confidentiality concerns, largely account for the lower e-mail response rate.

Table 8
Contacted Persons' Responses to Reminder Call, by Mode
(in percent)

	Mail	E-mail
Response to reminder call:		
Already returned	24.0	22.7
Will return	53.1	39.9
Refused, other	22.9	37.3
Total	100.0	100.0
Among those who refused, reasons given for nonreturn:		
Did not receive	3.9	8.0
Lost, deleted	3.7	9.1
Couldn't edit	0.0	2.7
No time	4.2	4.4
Confidentiality	1.4	1.2
Other, no reason	9.7	12.0
Total refused, other	22.9	37.3

A final source of data for evaluating the mode experiment comes from the *substantive responses* themselves. Given random assignment to mode, we would expect the distributions of key variables and the levels of item missing data to be similar. Table 9 contains item missing data rates by mode, for the 81 attitude items and 8 of the background items.

We see from Table 9 that the overall rates of missing data are low for both modes (on average less than 1 of the 81 attitude items missing per case). There are no significant ($p > .05$) differences in item missing data on the attitude items. Contrary to expectation, the mail mode has a significantly higher ($p < .01$) rate of missing data on the background measures. Inspection of the individual items suggests that several (*e.g.*, years of service, grade level, managerial and supervisory status, and race) are susceptible to higher missing data rates on the mail questionnaire. One possible explanation may be the differential effect of nonresponse – those who did make the effort to complete the e-mail questionnaire may have been more motivated to provide complete information. This again suggests that confidentiality may not have been a major factor in noncompletion.

Table 9
Item Missing Data Rates by Mode

	Mail	E-mail
81 attitude items	0.63	0.64
8 background items	0.24	0.16
(n)	(2,969)	(1,724)

We also examined the distributions of both demographic and substantive variables across mode. We assume that those who use computers more routinely in their work (*e.g.*, those in higher grades) would be more likely to return the e-mail questionnaire. We find significant differences ($p < .01$ in each case) in the distributions of respondents in terms of grade level, managerial and supervisory status. These results are presented in Table 10. Overall, the direction of the effect is as expected: higher status employees appear to be over-represented in e-mail. These differences are striking, and suggest differential access to, or use of, e-mail. We also find significant differences by race and gender ($p < .01$), with non-minorities and males being more likely to respond by e-mail than by mail. These results are also presented in Table 10.

Table 10
Distributions of Respondent Demographic Characteristics, by Mode (in percent)

	Mail	E-mail
Grade level		
Grades 1- 4	20.2	2.8
Grades 5 - 11	32.6	25.5
Grades 12 - 13	34.9	53.1
Grades 14+	12.4	18.7
Managerial status:		
Yes	14.7	22.6
No	85.3	77.4
Supervisory status:		
Yes	23.5	31.2
No	76.5	68.8
Gender:		
Male	40	47.6
Female	60	52.4
Race:		
White	77.3	82.8
Black	17.3	11.0
Other	5.4	6.2

In terms of substantive differences on the climate items, we assumed that nonrespondents may hold more negative attitudes toward their agency. Thus, with the higher nonresponse rate for e-mail, we expected more positive attitudes among those who did respond, relative to mail. We compared the mean scores between the two groups on each of the 13 organizational climate subscales, as well as the overall mean climate score. We found significant differences by mode on 5 of the 13 subscales, with mail having a higher (more positive) mean score on 3 of the 5, and e-mail on the remaining 2. Overall, mail respondents were more positive on 7 of the 13 subscales, and e-mail on 6. Thus, we find little support for our expectation on attitude differences. We also found no significant difference in the overall climate score, or in various other key single-item indicators such as satisfaction with the agency or employee morale. Thus, despite the differential nonresponse, the substantive responses of the mail and e-mail samples appear similar.

Finally, while we do not have a detailed cost breakdown for the two modes, we can offer a few observations on the cost implications of our study. The task of evaluating and testing e-mail software took over 150 hours of staff time, or almost 4 times what was budgeted. Printing and postage costs were $13,600 for mail and $0 for e-mail. Keying the completed mail questionnaires cost about $5,400 (about $1.81 per completed case), whereas managing the e-mail sample (including the clerical action mentioned earlier) cost about $3,000 (or $1.74 per completed case). The SRC staff handled over 900 incoming toll-free calls regarding the survey, most of which were technical questions about e-mail. Given the relatively large start-up costs, technical problems associated with e-mail, the high level of clerical action required, and the low response rate relative to mail, in this study we did not experience the cost savings expected from e-mail.

5. CONCLUSION

While e-mail offers potential savings in time and money over mail for organizational surveys, it seems clear that such benefits will not always be realized. Most other studies of mail versus e-mail have been conducted in relatively closed settings (*e.g.*, within one organization), thus minimizing the technical difficulties we experienced. Despite pretesting the survey in each of the agencies, we did not anticipate the problems caused by size of message limitations on certain platforms. These problems suggest that simply because every sample person has an e-mail address, does not mean that they will receive the survey or be able to respond in the manner intended. One advantage of mail in such cases is that it ensures a standard treatment for all sample persons. E-mail clearly offers a lot of promise, but the technical limitations need to be overcome before e-mail can be routinely used for surveys of large and diverse populations across multiple organizations.

ACKNOWLEDGMENTS

The students in the JPSM practicum played an important role in the design and conduct of this study. Financial support came from the National Science Foundation, through its support of the JPSM. This was supplemented by contributions from the participating agencies. Agency involvement was coordinated through an advisory council led by Richard Schuchardt of NASS (who served as the liaison to the practicum class), Lynda Carlson and Emilda Rivers of EIA, and Cynthia Clark and Nancy Bates of the Census Bureau. Without agency willingness to participate in this risky venture, the mode experiment could not have taken place. We also thank Skip Camp who, as SRC study manager, invested considerable time and effort in this project.

REFERENCES

Bachman, D., Elfrink, J., and Vazzana, G. (1996). Tracking the progress of e-mail vs. snail-mail. *Marketing Research*, 8 (2): 31-35.

Carlson, L.T., and Rivers, E.B. (1997). Origins of the organizational climate survey of federal statistical agencies. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Anaheim, CA, August.

Mehta, R., and Sivadas, E. (1995). Comparing response rates and response content in mail versus electronic mail surveys. *Journal of the Market Research Society*, 37 (4): 429-439.

Schuldt, B.A., and Totten, J.W. (1994). Electronic mail vs. mail survey response rates. *Marketing Research*, 6 (1): 36-44.

Sproull, L.S. (1986). Using electronic mail for data collection in organizational research. *Academy of Management Journal*, 29 (1): 159-169.

Tse, A.C.B. (1996). Comparing the response rate, response speed and response quality of two methods of sending out questionnaires: E-mail vs mail. Paper presented at the International Conference on Social Science Methodology, Essex, July.

SESSION I-4

Coverage Measurement in Censuses

DEVELOPING A ONE NUMBER CENSUS IN THE UNITED KINGDOM

I. Diamond, A. Teague, J. Brown, L. Buckner, R. Chambers and D. Thorogood[1]

ABSTRACT

As a result of lessons learnt from the 1991 Census, a research programme was set up to seek improvements in Census methodology. Underenumeration has been placed top of the agenda in this programme, and every effort is being made to achieve as high a coverage as possible. In recognition, however, that 100 per cent coverage will never be achieved, the One Number Census (ONC) project was established to measure the degree of underenumeration in the 2001 Census and, if possible, fully adjust the statistics from the Census for that undercount, so that all counts add to 'One Number'. This paper describes the background to the project and the methodological research currently being carried out.

KEY WORDS: Census; Underenumeration; Coverage survey.

1. INTRODUCTION

One of the major uses of Censuses in the UK is in providing a new base for the annual estimates of the population by age and sex. (Three Censuses are carried out in the UK, one in England and Wales, one in Scotland and one in Northern Ireland. The three Censuses are being planned together; they will be carried out on the same day using similar methodology although there will be some minor differences in the questions asked in the four countries.) This base needs to take into account the level of underenumeration in the Census. This has traditionally been measured by the use of a post-enumeration survey (PES) and through comparison with the estimate of the population based on the previous Census. Until the 1991 Census, there was close agreement between the adjusted Census count (Census + PES) and the estimate based on the previous Census. Moreover, the estimated level of underenumeration was relatively small (less than 1 per cent). In 1991, the level of underenumeration was much higher (2.2 per cent); underenumeration did not occur uniformly across all socio-demographic groups and parts of the country (for example, over 20 per cent for young males in inner cities); and there was a significant difference between the survey-based estimate and that based on the previous Census. This led to difficulties for both the Census Offices in rebasing the population estimates and Census users in interpreting Census counts.

The priority for the development of the 2001 Census is to ensure that the maximum possible coverage is achieved, and in particular that the differential nature of any underenumeration is minimised. To this end, the methodology for carrying out the Census is being re-assessed; to reduce the burden on the public and to use resources to their best effect.

Despite efforts to improve coverage, it is only realistic to expect there to be some degree of underenumeration. The One Number Census project aims to measure this level of underenumeration in the most acceptable way, to provide a much clearer link between the Census counts and the population estimates, and if possible to adjust all the Census counts (which means the database itself) for underenumeration. All counts will then add to 'One Number'. This has entailed a re-think of the design of the post-enumeration survey and how this should be integrated with other indices of the undercount, provided by administrative records and demographic analysis. The next section of the paper provides an overview of the methodology.

2. OVERVIEW OF THE ONC METHODOLOGY

The process outlined below represents current thinking on the possible ONC methodology. However, as it is a development project, the methodology described here may be subject to change before possible implementation. The ONC process can best be considered as consisting of four main stages, summarised below and illustrated in Figure 1. The first two stages will produce the best estimate of the population by age and sex at national and *subnational* (average 1 million population) levels. The third stage will produce estimates for lower levels of geography and for other characteristics of people and households. The final stage is to impute records for households that are estimated to have been missed and for people estimated to have been missed from counted households. This last stage would allow all published statistics based on the 2001 Census to aggregate to 'One Number'.

2.1 Stages 1 and 2 – National and Subnational Level Estimates

To estimate the population by age and sex at the subnational level, counts from the 2001 Census will be

[1] Ian Diamond, James Brown and Ray Chambers, Department of Social Statistics, University of Southampton; Andy Teague, Lisa Buckner and David Thorogood, Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hampshire, PO15 5RR.

adjusted for estimated net underenumeration using a post-enumeration survey to be known as the *Census Coverage Survey (CCS)*. Estimates of the undercount will be made either by using a regression estimator or a capture-recapture approach. This latter method assumes independence between the Census and the CCS. In other words, the probability of an individual being enumerated in the CCS is independent of the probability of being enumerated in the Census. It is inevitable that there will be some dependence and details of research investigating various levels of dependency are given in Brown *et al.* (1998). In Section 3 the regression-based approach is described.

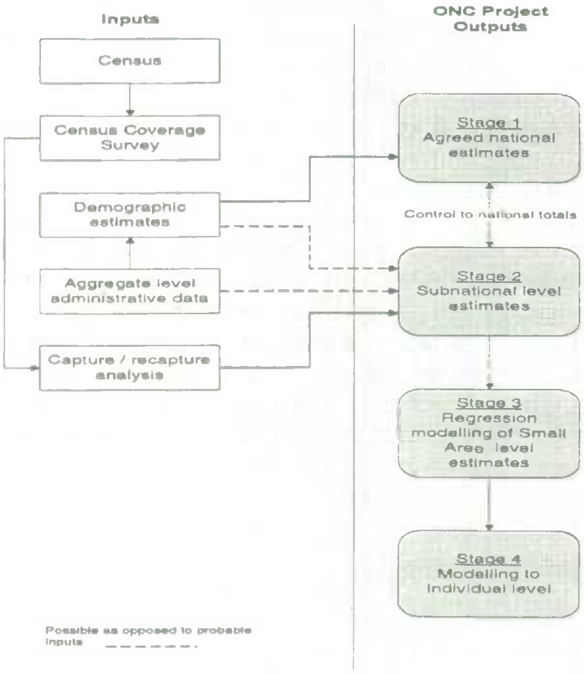

Inputs
ONC Project Outputs
Census
Census Coverage Survey
Demographic estimates
Aggregate level administrative data
Capture / recapture analysis
Stage 1 Agreed national estimates
Control to national totals
Stage 2 Subnational level estimates
Stage 3 Regression modelling of Small Area level estimates
Stage 4 Modelling to individual level
Possible as opposed to probable inputs

Figure 1. The stages of a One Number Census

The subnational level estimates will then be aggregated to produce a *national Census-based estimate* and compared with *the national demographic estimate* of the population (the estimate of the population rolled forward from the previous Census). Charlton *et al.* (1997) summarises work underway to optimise the methodology used to produce the demographic estimates.

A design for the CCS based on the re-enumeration of a sample of whole postcode units, stratified by county and an index based on predicted difficulty of enumeration, was successfully piloted in the Brent area of London following the 1997 Census Test. In the UK, unit postcodes cover around 12 addresses on average. A short questionnaire was used to collect information on characteristics believed to be associated with underenumeration. The simplicity of the questionnaire and the fact that sampling whole postcode units makes efficient use of interviewer time, makes a much larger sample size possible than was the case for the 1991 Census Validation Survey.

2.2 Stages 3 and 4 – Small Area Estimation and Imputation

The production of adjusted Census counts for small areas represent the final goals of the ONC process. Models developed for the postcodes sampled in the CCS linking the observed Census characteristics and the estimated missing population will be used to predict the number missed in non-sampled postcodes. These models will estimate the number of people missed in enumerated households and those in wholly missed households for each postcode. The precise method for creating individual records and allocating them to household units has not yet been developed but possible approaches are discussed in Section 5.

3. CENSUS COVERAGE SURVEY

This section outlines a proposed strategy for a Census Coverage Survey to be carried out after the 2001 Census. A model-based approach is adopted for the design and direct estimation. The aim is to estimate underenumeration at a subnational level by age and sex; and to allocate this underenumeration down to small areas. This will be necessary even if a full ONC is not implemented since estimates will have to be produced at the level that central government resources are allocated to local areas. This is currently Local Authority district level, however, estimates below this will probably be required to enable Local Authorities to allocate funding within their areas.

The proposal is for a postcode-unit based survey which will address coverage rather than both coverage and quality as in previous Censuses. There will be a re-enumeration of a sample of postcode units rather than households. This clustering of the sample permits a larger sample size. While, that does not necessarily improve the direct estimation, (due to clustering effects) it is important for estimating adjustments at lower levels. The sample of postcodes needs to be sufficiently large to estimate the total population by 24 age-sex groups, for each design level group. (Each design level group is either a single county or group of smaller contiguous counties.) At the design level, postcodes are stratified into groups by a 'Hard to Count' index and size. In this paper, 'hard to count' is defined in terms of characteristics found to be important after the 1991 Census by ONS and the Estimating With Confidence Project (Simpson *et al.* (1997). The problem is to estimate 24 age-sex totals such that each has a Relative Standard Error (RSE) of less than a certain percent at the design level.

In general, postcode level information, beyond number of addresses, is not known. This leads to a two-stage design, selecting enumeration districts as Primary Sampling Units (PSUs) and then sampling postcodes as Secondary Sampling Units (SSUs) within selected enumeration districts. Clustering from the two-stage design has cost advantages for a fixed number of postcodes but efficiency disadvantages when the characteristics of postcodes are positively correlated within enumeration districts.

3.1 Direct Estimation From the Census Coverage Survey

The quantities of interest are:

Z_{aidc} = 1991 adjusted Census count for age-sex group a of postcode i, in hard to count group d of county c.

X_{aidc} = 2001 unadjusted Census count.

Y_{aidc} = True 2001 count (given by the CCS for those postcodes in sample).

where:

$c = 1...C$ design level county groups in England & Wales.

$d = 1...D$ hard to count categories of postcodes.

$a = 1...24$ age-sex groups (0-4, 5-9, ..., 40-44, 45-79, 80-84, 85+).

$i = 1...N_{dc}$ postcodes in hard to count group d of county c of which n_{dc} are in the sample S, the rest are in the non-sample R.

For direct estimation from the CCS it is required that the total populations T_{ac} be estimated to a certain degree of accuracy. This is treated as 24 similar estimations within each design level group. For this reason the design and estimation for one age-sex by design level group is described below. The same model framework applies for all other age-sex groups and in the following the subscripts a and c are dropped.

3.2 CCS Design

A robust non-parametric model for stage one is a stratified homogeneous super-population model of enumeration districts with simple random sampling within each stratum. Within a design level group the enumeration districts are stratified by the hard to count index. This is important as, within the design group, undercount will depend on the characteristics of the PSUs. It also ensures that the CCS sample is spread across the full range of enumeration districts. Further stratification by size using the 1991 adjusted Census counts improves efficiency by reducing within stratum variance. Ideally one would like to use the 2001 counts but the CCS must be ready for the field directly after the Census so this is not possible. It is expected that the final design will use 1991 based estimates of the population in 2001.

Allowing for $h = 1...H_d$ size strata within each hard to count group, *and in this case using i for enumeration districts rather than postcodes*, the model for a given age-sex group within a design level group can be written as:

$$\left.\begin{aligned} E_\xi\{Y_{ihd}\} &= \mu_{hd} \\ \text{VAR}_\xi\{Y_{ihd}\} &= \sigma^2_{hd} \end{aligned}\right\} i \in h \text{ within } d$$

$$\text{Cov}_\xi\{Y_i, Y_j \mid X_i, X_j\} = 0 \text{ for all } i \neq j$$

It is possible to write-down a model for the second stage. This is more complicated due to the varying numbers of postcodes within enumeration districts, expected but unknown correlation of postcodes within enumeration districts, and the absence of readily available postcode level information on which to design. Therefore, the proposal is to have a constant second stage sample size and take a simple random sample of postcodes within chosen enumeration districts. This has the appeal of simplicity. Also, implicitly from the size stratification at stage one, and assuming the number of SSUs in a PSU is proportional to its size, this approach means that within a hard to count group each postcode has approximately the same inclusion probability.

3.3 The CCS Super-Population Model for Estimation

It is sensible to assume that the 2001 Census count and the CCS count within each postcode will be related. If this is not true then one really should be suspicious of one of the counts. Further, within sub-groups of postcodes a linear relationship may well be appropriate. This corresponds to a constant ratio (or adjustment factor) between the two counts with the possibility of a non-zero constant. This constant is needed as in some postcodes the Census can miss all people from a certain age-sex group. Given that we know from the 1991 Census that age and sex are crucial to undercount, as well as local characteristics, it is sensible to consider a model within age-sex groups for each hard to count by design level group where the hard to count index allows for different local characteristics. The simple regression model stratified by the hard to count index for an age-sex group is:

$$\left.\begin{aligned} E_\xi\{Y_{id} \mid X_{id}\} &= \alpha_d + \beta_d X_{id} \\ \text{VAR}_\xi\{Y_{id} \mid X_{id}\} &= \sigma^2_d \end{aligned}\right\} i \in d$$

$$\text{Cov}_\xi\{Y_i, Y_j \mid X_i, X_j\} = 0 \text{ for all } i \neq j$$

Substituting in the Ordinary Least Squares (OLS) estimators for α_d and β_d it is straightforward to show (Royall 1970) that under this model the Best Linear Unbiased Predictor (BLUP) for the total T is:

$$\hat{T}_\xi = \sum\nolimits_d \left\{T_{Sd} + \sum\nolimits_{Rd} \left(\hat{\alpha}_d + \hat{\beta}_d X_{id}\right)\right\} \tag{1}$$

where $\sum_{Rd}$ is the summation over all non-sample postcodes in hard to count index d and $\sum_d$ is the summation over all the hard to count groups. Strictly speaking the model is known to be wrong (the postcodes are correlated within enumeration districts), but the simple two stage model proposed by Scott and Holt (1982), which assumes independence between PSUs, is still reasonable. Under this model they state that the OLS approach remains unbiased and the loss of efficiency has a small upper bound when the residual correlation within clusters is small.

There is a model-based formula for estimating the variance of , $\hat{T}_\xi - T$, the estimation error, under the model. Unlike the estimator of the total, this is sensitive to mis-specification of the variance structure even when the design is *approximately* balanced with respect to the auxiliary variable (Royall and Cumberland 1978). In this strategy it is proposed that the, in general, conservative ultimate cluster variance estimator, a variant of the random groups approach, be used as the postcodes are clustered within enumeration districts. Once the variances are estimated an estimated RSE can be calculated for each age-sex group total. In general, when the regression model is appropriate, the estimator in (1) is more efficient than a simple stratum by stratum expansion estimator for a given sample size.

4. SIMULATION STUDY OF THE DESIGN AND ESTIMATION PROCEDURES

The detail in Section 3 only really examines the first stage of the design. The aim of the simulation is to examine the performance of the design when the second stage sample is taken. It is also necessary to see how appropriate and efficient the regression estimator is. Extra efficiency is also needed since in 2001 the design will be based on 1991 data.

Anonymised individual records, augmented by a Hard to Count index (HtC), from the 1991 Census for one complete district from each of six counties of England and Wales were used in the simulation. Each district is treated as a county in the design. The initial results presented here are for the largest county, which has 445,351 individuals within 171,265 households. It consists of 11,330 postcodes (141 with only 1 person and 46 with over 200 people) and 930 enumeration districts (5 have only 1 postcode, 1 has 40 postcodes, and the median is 14 postcodes). The distribution of enumeration districts by Hard To Count Index is given in Table 1.

The distribution in Table 1 is fairly even with respect to the index. This is a good test as it is necessary to avoid extremes, especially a situation where the very easy group dominates as this would tend to make the overall performance of the design too optimistic.

Table 1
Distribution of enumeration districts by hard to count index

Hard to Count index	Number of Enumeration
Very Easy	144
Easy	210
Medium	186
Hard	193
Very Hard	197

The first stage was to create a 'Census'. Each individual was given a fixed probability of being counted in a Census based on their age, sex, and enumeration district hard to count index. This was done by simple random sampling with replacement from the population of Estimating With Confidence enumeration district adjustment factors (Simpson *et al.* (1997). These are the best guess at small area underenumeration in the 1991 Census. To create a 'Census' a binomial trial was carried-out for each individual. This was done independently and certain rules were then applied to ensure that counted households had a sensible structure, *e.g.*, households with no adults were not allowed but one parent could go missing and the household would still be included. The CCS design procedure was based on an RSE of 2.5 percent to reflect the relatively smaller population of PSUs.

The design in Table 2 was fixed throughout the simulation and used to get a total sample of 85 enumeration districts. A fixed sample of four postcodes (or the number of postcodes in the enumeration district if less than four) was taken at the second stage. For each sample the totals for each age sex group were estimated, the variances calculated using the ultimate cluster variance estimator and estimated RSEs calculated. Ideally, it would be desirable to simulate one 'CCS' per 'Census' but for computational reasons 10 'CCSs' for each of 100 'Censuses' were used.

Table 2
Sample allocation for the first stage sample in County One

Index Group	Number of Enumeration Districts	Number of Size Strata	Sample Size	Outliers[b]
Very Hard	144	10	12	0
Hard	210	16	17	0
Medium	185	14	14	3
Easy	192	15	18	3
Very Easy	197	15	16	0
Outliers[a]	2	–	2	–
Total	930	70	79	6

[a] Enumeration districts classified as outliers based on their size.
[b] Enumeration districts classified as outliers by the clustering algorithm.

The mean total coverage across the 100 Censuses is 95 percent, the worst age sex group is males 20-24 with an average coverage of 89 percent. This is in general conservative for most counties compared to 1991. A few counties did do worse, particularly Inner and Outer London and those containing the large metropolitan districts (Heady *et al.* 1994). The key measure of performance of the procedure is the estimated RSEs. The results demonstrate that the procedure does well and in all cases the average estimated RSE is better than the RSE predicted by the model framework used in the design. This shows that on average the regression estimator is efficient enough. However, the standard errors do show that in most cases a significant percentage of CCSs still do worse than the design predicts and it cannot be guaranteed that the regression estimator will do better. Full details are found in Brown *et al.* (1998).

The initial conclusions are that the regression estimator is working well to recover any loss of efficiency due to the second stage design and multivariate stratification. However, the spread of RSEs is quite high. Another good point is that the ultimate cluster variance estimator is giving good coverage and 95 percent confidence intervals include the true value in at least 950 cases. Further work includes repeating the simulation for the other counties, investigating the effect of removing the second stage sample, and increasing the number of enumeration districts but only taking one postcode per enumeration district. It is expected that due to a positive correlation of postcodes within enumeration districts this last option will be more efficient, for a fixed number of postcodes, than the current procedure. It is known to also, in general, be more costly due to increased field costs as a result of interviewers having to travel further between postcode units.

4.1 Census Coverage Survey Pilot

The Brent Pilot Census Coverage Survey was carried out following the 1997 Census Test. The purpose of this survey was to assess the practicalities of undertaking a postcode based CCS. During the survey questionnaires were completed for 196 households containing 479 individuals. However, there was a wide variation in the contact and response levels between enumeration districts. Generally, and somewhat surprisingly, the higher response levels were achieved in the areas which were considered to be 'harder to enumerate'. The study concluded that it was possible to undertake a postcode based survey such as this but that it would be important to research further the optimal fieldwork strategy which would be short but intensive. Further details are given in Holland *et al.* (1997).

5. ADJUSTING INDIVIDUAL RECORDS – A FULL ONE NUMBER CENSUS

The ultimate aim of the project is a single individual level Census database fully adjusted for estimated under-enumeration. This requires a procedure that allows the imputation of missing people for both counted households and missed households. This section proposes a modelling approach to the problem. Note, however, that it is currently being tested empirically and so only initial results are available.

5.1 Situation after the Census and CCS

Let us assume that the CCS has taken place in a sample of postcodes within each design level group. Without loss of generality only one design group is considered. For those postcodes in the sample there are two listings of individuals, one from the Census and one from the CCS. These lists can, in principle, be matched to produce a single list of individuals containing *all* Census individuals with any *extras* from the CCS. This is a slightly different assumption to the one in Section 4 and recognises that the CCS will not find all the people that the Census does. The assumption is that no one is missed by both.

At the individual level one has:

(i) a matrix of socio-economic characteristics **X**
(age, sex, marital status, ethnicity, economic status)

(ii) a matrix of household characteristics **Z**
(tenure, building type, multiple-occupied, number of residents)

(iii) a vector of the household structure $\underline{S}$

Each individual i belongs to a household j within postcode k within enumeration district 1 of district m. The CCS does not contain all districts or postcodes so there is a prediction problem for the non-sampled postcodes. From the CCS direct estimation, demographic analysis, and capture recapture modelling there are *gold standard* age sex totals. The goal is to share the 'extra' people amongst the enumeration districts.

5.2 Multinomial Model for Small Area Adjustments

In relation to the assumption above, consider the possible categories of enumeration into which an individual can fall. Person is either counted, missed in a counted household, or missed in a missed household. This can be represented by the dependent variable y_{ijklm} where:

$y_{ijklm} = 0$ when individual i is counted in the Census (but not necessarily the CCS as well)

$y_{ijklm} = 1$ when individual i is missed in the Census and household j is counted (with respect to the CCS)

$y_{ijklm} = 2$ when individual i and household j are missed in the Census (with respect to the CCS)

and in general these outcomes will depend on the characteristics of the person, household, postcode, *etc*. Putting aside measurement error problems the following multilevel multinomial model can be fitted for the CCS sarnple postcodes:

$$\ln\left(\frac{\pi_{1ijlkm}}{\pi_{0ijlkm}}\right) = \alpha_1 + \underline{\beta_1'}\,\underline{X}_{1ijklm} + \underline{\gamma_1'}\,\underline{Z}_{1jklm} + \underline{\eta_1'}\,\underline{S}_{1jklm} +$$

$$\sigma_{1m} + \varsigma_{1lm} + \lambda_{1klm} + \upsilon_{1jklm} + \varepsilon_{1ijklm}$$

$$\ln\left(\frac{\pi_{2ijklm}}{\pi_{0ijlkm}}\right) = \alpha_2 + \underline{\beta_2'}\,\underline{X}_{2ijklm} + \underline{\gamma_2'}\,\underline{Z}_{2jklm} + \underline{\eta_2'}\,\underline{S}_{2jklm} +$$

$$\sigma_{2m} + \varsigma_{2lm} + \lambda_{2klm} + \upsilon_{2jklm} + \varepsilon_{2ijklm}$$

This is a standard random intercepts model and is important for small area estimation as this allows for extra heterogeneity between postcodes and enumeration districts. Without this all postcodes converge to the mean set by the model.

5.3 Prediction for non-Sampled Postcodes

As not all postcodes are in the sample, the first stage is to use $\underline{\hat{\beta}_1}$, $\underline{\hat{\beta}_2}$, $\underline{\hat{\gamma}_1}$, $\underline{\hat{\gamma}_2}$, $\underline{\hat{\eta}_1}$ and $\underline{\hat{\eta}_2}$ to get predicted probabilities for each of the different types of individuals and households in all areas. Again ignoring measurement error issues, this is straightforward for the fixed effects model but not for the multilevel model. For the latter case there is no estimate of higher level residuals. This is due to the independence assumption made in the multilevel framework. Ideally one would like to fit full spatial random effects. Spatial does not need to mean geographic. It may be more appropriate to 'borrow strength' from other areas based on distance measured in terms of demographic characteristics. This reflects the situation where, especially in cities, rich and poor live in contiguous areas. Computationally speaking this is currently extremely difficult. A proposal which is currently being considered is to fit the model in the independence framework and then use a spatial smoothing function to estimate residuals for non-sampled postcodes assuming the random effects are significant. This means that in principle for all areas one can estimate $\hat{\pi}_{0ijklm}$, $\hat{\pi}_{1ijklm}$, $\hat{\pi}_{2ijlkm}$.

5.4 Adjusting the Census

Let N_{ijklm} be the Census count of individuals with the set of characteristics given by *ijklm*. (*eg*., white 20-24 married employed male renting a detached house who is a member of a nuclear household of size 3 within postcode *k*.)
P (People of type *ijklm* are counted) $= \hat{\pi}_{0ijklm}$.

From this the number of people of type *ijklm* who are missed is given by:

$$N_{ijklm} \times \left(\frac{1}{\hat{\pi}_{0ijklm}} - 1 \right) = N_{ijklm} \times \left(1 - \frac{\hat{\pi}_{0ijklm}}{\hat{\pi}_{0ijklm}} \right) = \hat{N}_{ijklm}$$

The problem is now how to allocate these 'extra' people to already counted households or completely missed households. It can be shown that the number of missed people from counted households and the number of people missed from missed households are respectively:

$$N_{ijklm} \times \left(\frac{1 - \hat{\pi}_{0ijklm}}{\hat{\pi}_{0ijklm}} \right) \times \left(\frac{\hat{\pi}_{1ijklm}}{1 - \hat{\pi}_{0ijklm}} \right) = N_{1ijklm} \quad \text{and}$$

$$N_{ijklm} \times \left(\frac{1 - \hat{\pi}_{0ijklm}}{\hat{\pi}_{0ijklm}} \right) \times \left(\frac{\hat{\pi}_{2ijklm}}{1 - \hat{\pi}_{0ijklm}} \right) = N_{2ijklm}$$

where $\hat{N}_{ijklm} = N_{1ijklm} + N_{2ijklm}$ and the adjustments come directly from the multinomial model.

For the people from counted households (N_{1ijklm}), the task is to search the postcode for suitable donors based on the household characteristics, select a household using a random number generator or nearest fit criterion, and add the person. In certain cases the donor households will need a different structure to that of the new person. For example the donor household for a married man from a nuclear family is a single mother in the unadjusted Census.

For the people from the missed households, there will be a set of groups of people given by the different N_{2ijklm}. The task is then to fit the individuals back together as households. One possible way would be through a simulation which built-up households from available individuals. Another solution would be through an iterative proportional fitting algorithm where the N_{2ijklm} form marginal totals for types of individuals and the cells would be completed households.

5.5 Initial Results

Initial results using a simple fixed main effects model are promising and show that the method works. The results suggest that interacting sex with certain age groups (0-4, 20-34, 85+) will improve results for both males and females. In the reality of the One Number Census, one would expect to do even better by fitting random effects to account for additional small area variability, (random effects have not been fitted yet as the current simulation has no small area variability beyond enumeration districts belonging to the same hard to count group).

6. CONSULTATION

It is important that users of Census data have confidence in the figures produced from a ONC. Consultation will play a key role in the development and user acceptance of the methodology. Concern has been expressed by some that a full ONC will result in a significant delay to the availability of statistics from the Census. While there is likely to be some delay, preliminary discussions with major Census users – those responsible for allocating central government funds to local and health authorities – have indicated a willingness to wait for the ONC results provided it can be demonstrated that the product is possible and that any delay is not inordinate.

A Steering Committee, including representatives of the academic and local authority communities oversees the methodological development work. A senior representative of the Australian Bureau of Statistics is also a member of this committee. A major consultation paper will be issued early in 1998 to all interested parties. This will explain the ONC methodology as envisaged, in the context of the timing of the outputs and the marketing strategy for Census products. Feedback will be requested on this paper. There will be ongoing consultation with users through the Census Offices' user advisory groups.

The decision on the extent to which the ONC will be pursued will be taken at the end of 1998. This will be based on the likely quality of the resulting estimates from a full ONC and an assessment of the time and resources to do so.

REFERENCES

Brown, J.J., Buckner, L.J., Diamond, I.D., Chambers, R.L., and Teague, A.D. (1998). A Methodological Strategy for a One Number Census in the United Kingdom. To be submitted to *J.R.S.S.A.*

Charlton, J., Chappell, R., and Diamond, I.D. (1997). Demographic analyses in support of a One Number Census. To appear in *Proceedings: Symposium 97, New Directions in Surveys and Censuses*, Statistics Canada, November 1997.

Heady, P., Smith. S., and Avery, V. (1994). *1991 Census Validation Survey: coverage report*, London: HMSO.

Holland, F., Buckner, L.J., and Diamond, I.D. (1997). Report on the Brent Census Coverage Survey (CCS) Pilot. (Unpublished but available from the Office for National Statistics on request).

OPCS (1993). Rebasing the annual population estimates. *Population Trends*, 73, 27-31.

OPCS (1994). Undercoverage in Great Britain. *1991 Census User Guide 58*, London: HMSO.

Royall, R.M. (1970). On finite population sampling under certain linear regression models. *Biometrika*, 57, 377-387.

Royall, R.M., and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of American Statistical Association*, 73, 351-361.

Scott, A.J., and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of American Statistical Association*, 77, 848-854.

Simpson, S., Cossey, R., and Diamond, I. (1997). 1991 population estimates for areas smaller than districts. *Population Trends*, 90, 31-39.

SESSION I-5

Estimation and Variance meet the 21st Century: Computer Intensive Methods and Resampling

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

STATISTICAL ANALYSES OF COMPLEX SURVEY DATA USING WESVARPC SOFTWARE

K. Rust [1], J.M. Brick, D. Morganstein, T. Krenzke and P. James

ABSTRACT

The WesVarPC software calculates sampling errors for complex survey estimates, using replication methods (the jackknife and Balanced Repeated Replication), via a Windows interface. As well as providing estimates of standard errors for parameter estimates, the software can be used to conduct statistical tests of hypotheses for cross-classified data and linear and logistic regression models. There are a number of competing approaches in the literature to the calculation of test statistics. WesVarPC implements Rao-Scott first and second order corrections to Pearson's chi-square test of association in a two-way table. For linear and logistic regression, adjusted Wald-*F* statistics are used to test hypotheses about model parameters. We discuss the reasons for preferring these tests, as opposed to alternatives. We also indicate how some of the alternative approaches can be implemented simply using the softwarse capabilities, even though they are not directly supported.

KEY WORDS: Jackknife; Replication; Variance estimation.

1. INTRODUCTION

The conduct and analysis of large scale sample surveys generally combines the use of a complex sample design to obtain a sample of the population, coupled with an estimation procedure that reflects variations in the selection probabilities of the sample units, and often various adjustments to reduce nonsampling biases and sampling variance. The presence of these features means that in many cases inappropriate population inferences are obtained if one conducts analyses using statistical software that applies standards methods without incorporating the effects of these design and estimation features.

In recent years data analysts have become increasingly aware of the need to use care, and special procedures, when drawing inferences from complex survey data. There are two principal reasons for this. The first is the development of a literature that documents best practice, and clearly points out the pitfalls of failing to use appropriate procedures. As well as numerous articles in statistical journals, the books by Wolter (1985), Skinner, Holt and Smith (1989), and Lehtonen and Pahkinen (1996) provide detailed descriptions of the issues involved, and the techniques available, in analyzing survey data.

The second reason that valid inference from complex survey data is being practiced more commonly is that there is now available a range of statistical software that is explicitly designed to carry out these analyses appropriately. Such software is now widely available for use on PCs and work stations, as well as traditional main frame computers. Thus the analyst who wishes to draw valid statistical inferences has no excuse for ignoring appropriate statistical methods when analyzing complex survey data.

In this paper we will discuss one general approach to conducting analyses of complex survey data – replication – and one software package that utilizes this method – WesVarPC. After briefly outlining the replication approach, and the capabilities of WesVarPC, the paper concentrates on one aspect of the software. This is the choice of test statistics that are used in the software to make statistical inferences about multivariate relationships in the data. Specifically we discuss tests of association in two-way tables, and tests of model significance in linear and logistic regression models.

The literature cites a number of asymptotically equivalent test statistics that can be used in these cases. In this paper we discuss the rationale for choosing to implement some and not others in the WesVarPC software, and describe how the user can implement some of the other approaches, appropriately modifying input to and output from the software.

We note here that the technique of linearization provides an alternative approach to replication for the analysis of complex survey data. Literature that compares the methods of linearization and various forms of replication has shown that, while differences in results do exist, in many practical applications the results from the two approaches are extremely similar. We do not discuss linearization further here, since it is not implemented in WesVarPC. It is described in the three books mentioned above. There are software packages that use linearization, and there is other software that uses replication. Comparisons of the capabilities of these various PC software packages are given in Cohen (1997).

2. VARIANCE ESTIMATION USING REPLICATION TECHNIQUES

Replication methods, or resampling techniques, are a series of related approaches that utilize a very general, but

[1] Keith Rust, Westat, 1650 Research Blvd., Rockville, MD 20850, U.S.A.

computationally intensive, method for calculating estimates of sampling variance and deriving tests of statistical significance. The replication approach consists of estimating the variance of an estimate of a population parameter of interest by using a large number of varying subsamples (or by varying the sampling weights of the sample units) to calculate the statistic of interest. The variation among the resulting estimates is used to estimate the sampling variance of the initial, full-sample estimate.

There are three general replication approaches used to analyze complex survey data, each with a range of variants. These three methods are the jackknife, balanced repeated replication (or balanced half-samples) and the bootstrap. The methods differ principally in the procedures that they use to draw subsamples from the full data set. They each estimate the parameter of interest from each such subsample, and derive the sum of squared differences of these subsample estimates from the full sample estimate as the basic quantity used in estimating sampling variance (in some cases, the mean of the subsample estimates is used rather than the full sample estimate).

We denote the parameter of interest as X, and its full sample estimate as x. We denote each of the T replicate estimates that are formed in a given application of a replication procedure as $x_{(t)}$ for $t = 1, ..., T$. All replication procedures share the common feature that the variance of x can be estimated via a formula of the form

$$v(x) = \sum_{t=1}^{T} c_t.(x_{(t)} - x)^2. \qquad (1)$$

The WesVarPC software makes use of this feature; the different options in the software provide for different values of the constants c_t. These constants do not depend upon the parameter X or estimator x, but only upon the sample design and form of replication used. Thus a single set of constants c_t apply to all analyses of a given survey. In many cases, the c_t share a common single value (such as 1 or T^{-1}).

For discussions of the properties of replication variance estimators, see for example, Lehtonen and Pahkinen (1996), Rust and Rao (1996), and Wolter (1985).

3. THE WESVARPC SOFTWARE

This paper describes Version 2.1 of WesVarPC. The program produces sampling errors for functions of weighted totals specified by the user (*e.g.*, totals, means, percentages, differences, ratios, difference of ratios, log-odds ratios, *etc.*) and for quantiles. For two-way tables, WesVarPC performs a test of association (independence) using Rao-Scott corrected chi-square statistics (Rao and Scott 1984). It also computes sampling errors for linear and logistic regression model parameters and conducts tests on these parameters using Wald statistics.

3.1 PC Windows Environment

While its predecessor, WESVAR, was a SAS Proc running on a mainframe computer (IBM or VAX), WesVarPC was developed for operation in the Windows-PC environment. It is a stand-alone program written in the C language and requires only Windows (3.1 or 95). WesVarPC running on a Pentium PC produces results almost as quickly as WESVAR running as a SAS Proc on a mainframe.

The Windows environment facilitates development of programs that are easy to learn by those familiar with Windows. Unlike other statistical programs operated in a batch processing mode, WesVarPC users select variables by clicking in scrolling lists of variable names. The user's intent is conveyed by selecting choices from pull-down menus or by clicking on buttons to accept or cancel requests. Functions of variables and functions of table cells (such as log odds ratios) are specified in a calculator-like environment. Once a request is constructed, it is stored in a retrievable format, allowing subsequent recall and modification.

3.2 Capacity and Limitations

It is easy to enter survey data into WesVarPC because data files can be in several common formats: a plain ASCII text, PC-SAS or PC SPSS file. The program will read files with hundreds of variables and thousands of records. WesVarPC has been tested with files of 200,000 records and 80 replicates. Computation time for tables was no more than fifteen minutes on a Pentium-class PC.

Since WesVarPC uses a replication method for computing sampling errors, replicate weights are needed. Two options are available to the user: (1) provide the weights as a part of the data file or (2) use WesVarPC to create the replicate weights. The program's capability for creating weights, while somewhat limited, is adequate for many needs. This is discussed in more detail later.

WesVarPC reads the user's survey data file and creates an independent specially formatted version of that file. This initial step is performed once and is the most time-consuming part of using the program. Once created, this special WesVarPC version of the file can be processed very quickly as it contains important summary data that facilitate computations.

By using replication methods, WesVarPC offers several advantages over other programs that rely upon linearization. Missing data are handled in a uniform way. Subpopulations can be analyzed easily. There are no special routines that must be called or special care required to subset the data file. If analysis focuses entirely on a subgroup, then those survey responses can be placed in a special "subpop" file with a single command and analyzed more rapidly due to the smaller size of the "subpop" file.

As mentioned earlier, WesVarPC will either read previously created replicate weights found on the input survey data file, or create them using information provided by the user. To create weights, the program merely requires the identification of two variables on the file indicating each record's stratum and PSU membership. It can create up to 256 replicates. In addition to forming base replicate weights, WesVarPC can perform poststratification and, by repeated poststratification, raking (iterative proportional fitting). The user supplies one or more sets of control counts, within each set one count for each stratum. The full sample and replicate weights are adjusted to match those totals.

WesVarPC can create replicate weights for regular BRR with a two PSU per stratum design, the jackknife with a two PSU per stratum design (JK2), and an unstratified jackknife (JK1). But if the user provides a file that already contains replicate weights, WesVarPC can accommodate virtually any replication scheme. If the appropriate replicate weights are provided by the user, then the only requirement to calculate variances appropriately are the values of the constants c_i described in Section 2. These can be input to the program by the user. Thus WesVarPC can be used to calculate variances using the bootstrap, Fay's BRR (see Judkins 1990), and the jackknife with more than two units per stratum, for example.

3.3 User Interface

WesVarPC actions are selected via pull-down menu items. The desired variables are selected from a scrolling list by clicking on their names. Functions of variables are specified by clicking on variable names and using a simple calculator display to select the desired function. Sampling errors for basic statistics such as totals, averages and percentages, are obtained by simply clicking on the variable, specifying an average function or defining the table within which the percentages are wanted. WesVarPC employs a graphical user interface (GUI) to reduce the amount of learning needed to use the program.

WesVarPC allows the user to specify a multi-way table of up to eight dimensions within which statistics can be calculated. For example, the sampling error of average income can be computed within a three-way table defined by region of the country, age (in categories) and gender of the respondent. The user simply clicks on the three variable names that define the table, then clicks on the average function and selects the income variable from the variables pick list. Functions of cell computations can also be requested easily. For example, the sampling error of a log-odds ratio can be defined by naming the four cells of the table and specifying that a log is to be taken of the odds ratio for the estimated cells counts.

While the user is specifying a variety of computational requests by selecting analysis variables, defining functions and tables, WesVarPC is recording the request in a data file. This permanent request file can be re-opened at a later time, modified and re-submitted. Requests for sampling errors of specific variables can be added or subtracted easily. By double clicking on variables that were previously selected, they are removed. Similarly, new tables can be added to the request and old ones deleted, simply by clicking on the specified table. However, for users who require it, WesVarPC provides a batch request method to produce large numbers of tables and models. WesVarPC offers the benefits of both an easy-to-use graphical interface and, alternatively, a batch system for production runs of tables without the need for the user to learn a special purpose language.

The program's documentation is provided in two formats: an easy-to-read manual with an appendix containing technical details including explanatory formulas; and internal help screens. The manual is written for a researcher who has at most a limited understanding of the theory behind sampling error computations. It assumes very little knowledge (outside of using Windows) on the part of the user. The manual contains example screens of virtually every step the user will encounter when running the program. Every screen includes a Help button for accessing an explanation of the currently available options.

4. TEST STATISTICS IN WESVARPC

Although much of the emphasis in WesVarPC is on the estimation of standard errors for estimates of population parameters, and derivation of confidence intervals for the population parameters, WesVarPC does have the capability to implement certain specific multivariate hypothesis testing procedures. In the TABLES portion of the software, chi-square tests of association in two-way tables are provided as options. In the REGRESSION portion, multivariate tests that simultaneously test the statistical significance of several parameters are available, both for linear regression and logistic regression.

In each of these testing situations, the available literature provides several asymptotically equivalent alternative approaches to conducting such tests. WesVarPC implements only some of these. The following sections discuss the options available in the literature, and documents why the developers of WesVarPC have chosen to implement some but not others. As will be seen, these choices are a mixture of considerations of the theoretical and especially the empirical research findings that are presented in the literature concerning these alternatives, and practical considerations in implementing the methods in conjunction with a replication-based approach.

Before proceeding to those discussions, however, we point out that WesVarPC has a somewhat more general hypothesis testing capability than might at first be recognized. Many multivariate hypotheses of interest can be implemented by adopting the following general approach for testing a p dimensional hypothesis.

1. Identify p population parameters, such that the truth of the null hypotheses corresponds to all p parameters having values of zero.
2. Obtain appropriate formulae for estimating these p population parameters using survey data.
3. Use the COMPUTE and FUNCTION capabilities of WesVarPC to estimate each of the p parameters, and obtain sampling error estimates for these.
4. Use the Bonferroni (or other similar multiple comparison testing) method, in conjunction with WesVarPC's capability to estimate confidence intervals, to conduct p t-tests, one for each parameter. If the null hypothesis is rejected in any case, then the null hypothesis of the original p dimensional hypothesis test is rejected.

Although in general one might expect that a reliance on methods for conducting simultaneous one dimensional hypothesis tests might lead to a substantial loss of power in comparison to the appropriate multivariate testing procedures, evidence produced in recent years suggests that this is not necessarily the case. This is because in many complex survey applications, there are limited degrees of freedom available for variance estimation. In the context of testing hypotheses about linear models, Korn and Graubard (1990) have shown that this approach using multiple comparison procedures can be superior to traditional methods in some cases. Similar conclusions have been reached by Thomas *et al.* (1996) in the case of testing association in two-way tables.

The following sections contain examples of this approach in a multi parameter setting. But this approach can also be useful in a one parameter case. Consider the case of the Mantel-Haenszel common odds ratio across subpopulations (which may or may not constitute the survey strata). Biostatistical texts present a one degree of freedom chi-square test, to test the hypothesis that the common odds ratio is 1.0. WesVarPC is able to calculate (through the use of the COMPUTE feature) an unbiased estimate of the Mantel-Haenszel common odds ratio, together with an estimate of its standard error (or perhaps the logarithm of the common odds ratio, as using this statistic may give better small sample properties for the test). This can then be converted to a t test statistic, used to test the hypothesis that the common odds ratio is equal to 1.0 (or in fact any other prespecified value). This test has the advantage over the traditional chi-squared test in that it can account for the limited degrees of freedom available for variance estimation, through an appropriate choice of the degrees of freedom for the t test. Further, this approach is readily extended to situations such as comparing the size of the common odds ratio in two or more different domains, using a multiple comparisons approach in the case of three or more domains.

5. TESTS OF ASSOCIATION IN TWO-WAY TABLES

Thomas *et al.* (1996) have identified approximately 27 different procedures for testing association in a two-way $r \times c$ table constructed from complex survey data. They have conducted an evaluation of the properties of each approach. Lehtonen and Pahkinen (1996, Section 7.5) also suggest a set of approaches that Thomas *et al.* did not consider, based on the Neyman chi-square statistic (several of the approaches detailed in Thomas *et al.* are based the Pearson chi-square statistic). Yet WesVarPC offers the user only two such test statistics. In this section, we document the rationale for including the statistics that are provided, and also indicate that many of the other test statistics can readily be derived from WesVarPC output, with minimal additional calculation.

WesVarPC provides the user with a Pearson chi-square statistic and associated degrees of freedom and significance level, a Rao-Scott first order correction to Pearson's chi-square (with degrees of freedom and significance level) and a second order Rao-Scott correction, and associated degrees of freedom and significance level. The details of the calculations of these statistics are provided in the WesVarPC Version 2.1 documentation (Brick *et al.* 1997), as well as Rao and Scott (1984) and Lehtonen and Pahkinen (1996, Section 7.5), among other sources.

The two Rao-Scott corrections are included because, when first implemented in WESVAR (the mainframe SAS Proc) in the late 1980's these were considered to be methods with the best inference properties. In particular, they appear to have superior small sample properties to the alternative Wald statistics. The work by Thomas *et al.* (1996) indicates that these methods still retain desirable properties, but there are serious competitors.

The original motivation for offering both first and second order adjustments was two-fold. The first was that there was relatively little practical experience with using the second-order correction method, and users were not familiar with the concept of having non-integer degrees of freedom. The second was that the second order correction is much more computationally intensive than the first order correction (which is essentially derived as a by-product of producing standard errors for the estimates in the table cells). When conducting multiple analyses on main frame computers this was an important consideration. Evaluations of the properties of the two approaches, and the use of PCs (especially Pentiums) suggests that WesVarPC users should primarily consider the second-order Rao-Scott adjusted chi-square.

Krenzke (1997) has conducted a systematic evaluation of the capability of users to carry out each of the different methods identified by Thomas *et al.* either directly from the software output, or using simple manipulations of the output, such as could be carried out on a portable calculator or using spreadsheet software. Krenzke found that WesVarPC users can currently implement five additional approaches. Those that WesVarPC cannot accommodate principally require direct calculation of the eigenvalues of the design effect matrix. WesVarPC does not calculate these eigenvalues. It implements the Rao-Scott corrections (which are functions of these eigenvalues) using the linearization formulae derived by Rao and Scott (even though the variances and covariances themselves are calculated via replication).

There are several approaches that can readily be implemented using WesVarPC, which appear, based on the work of Thomas *et al.*, to provide possibly attractive alternatives to Rao-Scott corrected chi-square statistics. Three of these approaches involve converting the chi-square statistic to an F statistic, and thus endeavoring to account for the limited degrees of freedom available for estimating the variance-covariance matrix of proportion under a complex design. Thomas *et al.* strongly recommend an F statistic adjustment to the second order Rao-Scott chi-square. This involves dividing the Rao-Scott adjusted chi-square test statistic by its degrees of freedom, and comparing this to an F distribution with this same degrees of freedom in the numerator, and using for the denominator degrees of freedom an estimate of the degrees of freedom for the variance-covariance matrix. The major difficulty in applying this by manipulating the output from WesVarPC is finding the critical value for an F distribution with non-integer numerator degrees of freedom.

6. TESTS OF MULTIPLE PARAMETERS IN LINEAR MODELS

WesVarPC provides the capability of fitting multiple regression and logistic regression models, under the REGRESSION feature. One component of this is the ability to simultaneously test several hypotheses concerning the parameters in the model. A most commonly used test is one that tests whether all coefficients other than the intercept are equal to zero. Users also commonly wish to test whether a subset of the coefficients are equal to zero, or are equal to each other.

In conducting such tests using complex survey data, Shah *et al.* (1993) identify five related test statistics that can be used for this purpose. Each is based on calculating a Wald chi-square statistic, using a variance-covariance matrix estimated so as to take account of the complex design. These five statistics are:

1. The Wald chi-square statistic itself (denoted as Q), with critical value based on degrees of freedom given by the dimension of the null hypothesis. Thus if the test is that p regression coefficients are all equal to zero, the statistic has an asymptotic central chi-square distribution with p degrees of freedom if the null hypothesis is true.
2. A Wald F statistic, calculated as $F_W = Q/p$. This is evaluated against a F distribution with numerator degrees of freedom of p, and denominator degrees of freedom equal to the assumed degrees of freedom available for estimating the variance-covariance matrix of the coefficients, d (often taken as the number of replicates used in the jackknife or BRR procedures).
3. An Adjusted Wald F statistic, proposed by Folsom (1974). This statistic is somewhat more conservative than the Wald F, and accounts for the fact that some of the degrees of freedom for variance estimation are "used up" in estimating the parameters to be tested, so to speak. However, in the usual linear regression framework this derivation follows analytically from the fact that the variance-covariance matrix is estimated using the residual terms. Since this is not the approach used to estimate the variance-covariance matrix of the parameters in the complex survey case, whether the adjustment leads to improvement is more of an empirical question.

   The adjusted Wald F is calculated as $F_{AW} = Q.(d - p + 1)/d.p$, and is tested against the critical value for an F distribution with p and $(d - p + 1)$ degrees of freedom, where d is the assumed number of degrees of freedom for the variance-covariance matrix. Graubard and Korn (1993) provide evidence that this test performs well when d is large, relative to p.
4. Satterthwaite Adjusted Chi-square. This approach adjusts the Wald chi-square in an analogous fashion as the second-order Rao-Scott adjustment to the Pearson chi-square, discussed in Section 4. If e denotes the estimated mean eigenvalue of the generalized design effect matrix, and a is the estimated coefficient of variation of these eigenvalues, then the Satterthwaite adjusted chi-square is given as $S = Q^*/e(1 + a^2)$, where Q^* denotes the Wald chi-square, calculated ignoring the complex design. The value of S is then compared to a chi-square distribution with $p/(1 + a^2)$ degrees of freedom.
5. Satterthwaite Adjusted F statistic. This statistic is a logical extension of combining the approaches 2 and 4 above. That is, the Satterthwaite adjusted chi-square is divided by its degrees of freedom, and the result compared with an F statistic. Thus $F_S = S.(1 + a^2)/p = Q^*/(p.e)$ is compared to the critical value of an F distribution with $p/(1 + a^2)$ and d degrees of freedom.

The WesVarPC procedure gives just one of these five options. As S and F_S require the calculation of the variance covariance matrix ignoring the complex design, in addition to estimating it incorporating the complex design, then these two statistics have not been included in WesVarPC. Also, the F_S statistic has not received any extensive systematic evaluation in the literature. If future research suggests that either S or F_S has substantially superior small sample properties to the other three methods, then serious consideration will be given to implementing them in future editions of the software. Among the other three procedures, the adjusted Wald F statistic appears empirically to have the most desirable properties, and so it is included in WesVarPC. However, the user can easily obtain the necessary quantities from the output to calculate F_W and Q if so desired, since

$$\begin{gathered} F_W = F_{AW} \cdot d/(d - p + 1), \text{ and} \\ Q = F_{AW} \cdot d/(d - p + 1). \end{gathered} \tag{2}$$

In fact, if the user specifies the number of degrees of freedom available for variance estimation to be very large in comparison to the number of model parameters, then the F_{AW} statistic reverts to F_W.

All of these proposed methods tend to break down (*i.e.*, they give inappropriate rates of rejection when the null hypothesis is true) when the number of parameters to be tested is relatively large compared to the number of degrees of freedom for variance estimation; that is, if p is large relative to d. F_{AW} is not even defined in the case that p is larger than d. In circumstances where there are few degrees of freedom for variance estimation (most readily characterized in replication methods by the existence of only a small number of replicates), Korn and Graubard (1990) suggest using a Bonferroni approach to hypothesis testing. This is carried out by testing each parameter in turn using a t test with d degrees of freedom, and a significance level of α/p where α is the desired overall level of significance. If no single parameter achieves significance, the null hypothesis is not rejected. This can easily be implemented in WesVarPC, since the output provides the appropriate t test for each individual parameter, together with its significance level, and the analyst can then use the desired significance level for testing.

7. CONCLUSION

Replication methods, as implemented through WesVarPC, provide a broad range of capabilities for conducting multivariate testing procedures, as well as providing estimates of standard errors and confidence intervals for parameters. As future research refines the knowledge of best practice in conducting such tests, it seems likely that it will be straightforward to implement any future modifications through the replication procedures, and hence they will be incorporated in future versions of WesVarPC.

REFERENCES

Brick, J.M., Broene, P., James, P., and Severynse, J. (1997). A User's Guide to WesVarPC. Version 2.1. Westat: Rockville, MD.

Cohen, S.B. (1997). An evaluation of altemative PC-based software packages developed for the analysis of complex survey rata. *American Statistician*, 51(3), 285-292.

Folsom, R.E. (1974). National Assessment Approach to Sampling Error Estimation. Sampling error monograph prepared for the National Assessment of Educational Progress, Denver, CO.

Graubard, B.I., and Korn, E.L. (1993). Hypothesis testing with complex survey data: The use of classical quadratic tests statistics with particular reference to regression problems. *Journal of the American Statistical Association*, 88, 629-641.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

Korn, E.L., and Graubard, B.I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t*-statistics. *American Statistician*, 44(4), 270-276.

Krenzke, T. (1997). WesVarPC Capabilities for Tests of Independence on Two-Way Tables. Westat working paper.

Lehtonen, R., and Pahkinen, E.J. (1996). *Practical Methods for Design and Analysis of Complex Surveys*. Revised edition. New York: John Wiley.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Shah, B.V., Folsom, R.E., LaVange, L.M., Wheeless, S.C., Boyle, K.E., and Williams, R.L. (1993). Statistical Methods and Mathematical Algorithms Used in SUDAAN. Research Triangle Institute: Research Triangle Park, NC.

Skinner, C.J., Holt, D., and Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*. New York: John Wiley.

Thomas, D.R., Singh, A.C., and Roberts, G.R. (1996). Tests of independence on two-way tables under cluster sampling: an evaluation. *International Statistical Review*, 64(3), 295-311.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

VARIANCE ESTIMATION FOR PUBLIC USE MICRODATA FILES

W. Yung[1]

ABSTRACT

Many of Statistics Canada's surveys produce a Public Use Microdata File (PUMF) which is made available to analysts wishing to perform their own analyses. In addition to being able to produce simple descriptive statistics such as means and totals, enough information is included to perform more complex analyses such as linear or logistic regression. As part of disclosure avoidance procedures, design information such as stratum or cluster identifiers are not included on the PUMF. In the absence of this design information, users of the PUMF are unable to calculate correct variance estimators. Currently, users are informed of sampling variability by means of Approximate Sampling Variability Tables, which are valid for totals, ratios and proportions for categorical variables only. In this paper, we propose the method of bootstrap variance estimation as a solution to the problem of calculating correct design-based variance estimators for PUMF's. The bootstrap method will provide correct variance estimators for means and totals as well as for complex statistics such as estimated regression coefficients. Modifications to the commonly used bootstrap necessary to ensure confidentiality will be presented. The proposed method will be illustrated and compared to the current method using data from one of Statistics Canada's surveys.

KEY WORDS: Bootstrap; Confidentiality; Public use files; Variance estimation.

1. INTRODUCTION

Many of Statistics Canada's surveys produce a Public Use Microdata File (PUMF), which is made available to analysts wishing to perform their own analyses of Statistics Canada data. On these micro-data files, each record represents a sampled element (business, household, *etc*...) and includes a weight which usually incorporates adjustments for nonresponse and benchmarking. However, as part of disclosure avoidance procedures, design information such as stratum or cluster identifiers are not normally included on the PUMF. In the absence of this design information, users of Statistics Canada's PUMF's are unable to calculate valid design-based variance estimators. Currently, users are informed of sampling variability by means of *Approximate Sampling Variability Tables*. These tables give an approximate coefficient of variation (CV) for estimates of totals, ratios and proportions for categorical variables. Unfortunately, these tables cannot be used to obtain CV's for continuous variables or for complex statistics such as estimated regression coefficients. As well, this approach, in use since the 1970's, is now felt to be unsatisfactory for practical and statistical reasons.

In this paper, the use of the bootstrap method is proposed as a solution to the problem of producing valid variance estimates from PUMF's while still respecting confidentiality constraints. The bootstrap method can be used to calculate variance estimates for totals, ratios and proportions for categorical and continuous variables, as well as for more complex statistics such as regression coefficients. In section 2, the construction of the approximate sampling variability tables currently in use and the consequences of their construction are discussed. Alternative methods investigated at Statistics Canada are presented in section 3 while the proposed bootstrap method is described in section 4. Comparisons of the proposed method with the approximate sampling variability table method, using data from Statistics Canada's National Population Health Survey, are shown in section 5.

2. APPROXIMATE SAMPLING VARIABILITY TABLES

Approximate sampling variability tables, or CV look-up tables, have been used for many years as a means of informing microdata file users of sampling variability. Typically, these tables are produced at both the national and provincial levels and occasionally at sub-provincial levels. For each table, a set of approximately 30 key categorical variables is identified and exact variances are calculated for each response category cross-classified by age groups and sex. As well, variances are calculated under a simple random sample design and design effects (DEFF's) are obtained for each combination of response category, age group and sex. The 75th percentile of these DEFF's is then used as a representative DEFF for use in preparing the CV look-up tables. Use of the 75th percentile means that the estimated CV will be an overestimate 75% of the time and an underestimate the remainder of the time.

Users of Statistics Canada's PUMF's have expressed a wide range of views concerning the use of the CV look-up tables. Unsophisticated users do not understand how to use the tables and, as a result, tend not to use them. Others want only an easy procedure to determine the releasability of an estimate: releasable, qualified, or not releasable. Sophisticated users find the tables neither detailed enough nor adequate for complex analyses such as linear or logistic

[1] Wesley Yung, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

regression analyses. Still others find them burdensome because it is a manual procedure and they have many estimates for which they require CV's. In summary, it appears that there are two distinct groups of users:

1. Basic analysts for whom the CV look-up tables are appropriate (if and when they use them).
2. Sophisticated analysts who find the CV look-up tables burdensome and/or inadequate.

For the first group of users, it would be desirable to find a more automated method to obtain the approximate CV's, while for the second group a method for the analyst to calculate a valid design based variance estimator is desired.

3. ALTERNATIVE METHODS

At Statistics Canada, two alternative methods have been investigated as solutions to the PUMF variance estimation problem. The first approach utilizes the Generalized Variance Function (GVF) approach of Wolter (1985) while the second approach uses the jackknife variance estimator for a stratified multi-stage design in conjunction with collapsing of design strata and clusters.

3.1 Generalized Variance Functions

The method of generalized variance functions uses a mathematical model to describe the squared CV of a survey estimator and its expectation. Possible models include

$$CV^2 = \alpha + \beta/X$$

and

$$CV^2 = (\alpha + \beta X + \gamma X^2)^{-1}.$$

Wolter (1985) notes that there is very little theoretical justification for any of the models given above. For more on the choice of models, the reader is referred to Wolter (1985). To utilize the GVF, one would calculate many estimates of survey variables, $\hat{X}_i$ along with their corresponding CV's and then calculate $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ using ordinary or weighted least squares. Assuming the first model was chosen, an estimate of the CV of a survey statistic $\hat{Y}$ is then obtained as

$$\hat{CV}^2 = \hat{\alpha} + \hat{\beta}/\hat{Y}.$$

Although the GVF method appears to be a possible replacement for the CV look-up tables, it still suffers from some of the same problems experienced by the look-up tables. For the basic analysts, the GVF method provides an easy and somewhat automated method of obtaining approximate CV's. Unfortunately, attempts to develop GVF techniques for continuous variables have been largely unsuccessful. As well, the GVF method is not valid for more complex statistics such as regression coefficients. Even with these short comings, the GVF method is in use for the U.S. Bureau of the Census' Survey of Income and Program Participation (SIPP) and has been investigated for Statistics Canada's Survey of Labour and Income Dynamics (SLID).

3.2 Collapsing

For use with Statistics Canada's National Population Health Survey (NPHS), Mayda *et al.* (1996) proposed using the collapsing method of Rust (1986) to create 'super-strata' and 'super-clusters' and then applying the usual jackknife variance estimator on the super-strata and super-clusters. Following Rust (1986), design strata are collapsed to form super-strata and then the original clusters are collapsed within the super-strata. The clusters are collapsed in such a way that the super-clusters contain original clusters from the same design strata. The super-strata and super-cluster identifiers are then included on the PUMF, thus allowing analysts to use the jackknife variance estimator. This method is illustrated in Mayda et al. using data from the NPHS. Although results from their empirical study are encouraging, one should take care when collapsing strata and clusters within strata, as Valliant (1995) has shown that under certain conditions the balanced repeated replication (BRR) variance estimator can become inconsistent when strata are collapsed. It is unclear at this point whether the inconsistency property of the BRR extends to the jackknife variance estimator, but the asymptotic equivalent of the BRR and the jackknife variance estimators, as shown in Rao and Wu (1988), indicates that the jackknife variance estimator may also suffer due to collapsing.

4. BOOTSTRAP VARIANCE ESTIMATION

The bootstrap variance estimation method for the *iid* case has been extensively studied, see Efron (1982). Rao and Wu (1988) provided an extension to stratified multi-stage designs but covered only smooth statistics $\hat{\theta} = g(\hat{Y})$. The Rao-Wu bootstrap was extended by Rao, Wu and Yue (1992) to include non-smooth statistics as well as smooth statistics. The design considered by Rao, Wu and Yue, and in this paper, assumes L design strata with N_h clusters in the h-th stratum. Within the h-th stratum, $n_h \geq 2$ clusters are selected and further subsampling within selected clusters is performed according to some probability sampling design. Although the subsampling is not specified, it is assumed that there is unbiased estimation of cluster totals, Y_{hi}, h=1, ... , L; i =1,... , n_h. Based on the survey design, design weights, w_{hik}, associated with the *(hik)*-th sampled element are obtained. Also associated with the *(hik)*-th sampled element is the variable of interest, y_{hik}. An estimator of the total Y is then given by

$$\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik} \qquad (4.1)$$

where s denotes the sampled elements. The design weights are often subjected to adjustments such as poststratification or generalized regression to ensure consistency to known population totals. For example, suppose that each element in the population belongs to a poststratum that can cut across the design strata. Using prescript notation to denote poststrata, the total number of elements in the c-th poststratum is ${}_cM$ and is assumed to be known. Letting ${}_cw_{hik}$ represent the poststratified or final weight defined by

$${}_cw_{hik} = \frac{{}_cM}{{}_c\hat{M}} w_{hik},$$

where ${}_c\hat{M} = \sum_{(hik)\in s} w_{hik}\, {}_c\delta_{hik}$ and ${}_c\delta_{hik}$ is the poststratum indicator variable, the poststratified estimator is defined as

$$\hat{Y}_{ps} = \sum_{c} \sum_{(hik)\in s} {}_cw_{hik}\, y_{hik}\, {}_c\delta_{hik}.$$

To calculate a bootstrap variance estimator for $\hat{\theta} = g(\hat{Y})$, where $\hat{Y}$ is given by equation (4.1) and g is a known function, the Rao-Wu-Yue method proceeds as follows: (note that the poststratified estimator can be expressed in this form)

(i) Independently, in each stratum, select a simple random sample of m_h clusters with replacement from the n_h sample clusters.

(ii) Let m^*_{hi} be the number of times the *(hi)*-th cluster is selected ($\sum_i m^*_{hi} = m_h$). Define the bootstrap weights as

$$w^*_{hik} = \left[1 - \left(\frac{m_h}{n_h-1}\right)^{1/2} + \left(\frac{m_h}{n_h-1}\right)^{1/2} \frac{n_h}{m_h} m^*_{hi}\right] w_{hik}. \quad (4.2)$$

If the size of the simple random sample, m_h, is chosen to be less than or equal to n_h-1, then the bootstrap weights, w^*_{hik}, will all be positive.

(iii) To obtain the final bootstrap weight, perform the same weight adjustment with the design weights, w_{hik}, replaced by the bootstrap weights, w^*_{hik}. For example, the final bootstrap weight for the poststratified estimator is

$${}_c\tilde{w}^*_{hik} = \frac{{}_cM}{{}_c\hat{M}^*} w^*_{hik}$$

where ${}_c\hat{M}^* = \sum_{(hik)\in s} w^*_{hik}\, {}_c\delta_{hik}$.

(iv) Calculate $\hat{\theta}^*$, the bootstrap estimator of θ, using the final bootstrap weights, $\tilde{w}^*_{hik}$, in the formula for $\hat{\theta}$.

(v) Independently replicate steps (i) to (iv) a large number of times, B, and calculate the corresponding estimates, $\hat{\theta}^*_{(1)}, \ldots, \hat{\theta}^*_{(B)}$.

The bootstrap variance estimator for $\hat{\theta}$ is then given by

$$v_B(\hat{\theta}) = \frac{1}{B} \sum_b \left(\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(\cdot)}\right)^2$$

where $\hat{\theta}^*_{(\cdot)} = (1/B)\sum_b \hat{\theta}^*_{(b)}$.

A commonly used value for m_h is n_h - 1 in which case equation (4.2) reduces to

$$w^*_{hik} = \frac{n_h}{n_h - 1} m^*_{hi}\, w_{hik} \quad (4.3)$$

If a sampled element is in a cluster that has not been selected in a particular bootstrap sample, then $m^*_{hi} = 0$ and the bootstrap weight is equal to zero. That is, all sampled elements in the cluster have bootstrap weights equal to zero and in the case of multiplicative weight adjustments (*e.g.*, poststratification or regression), will also have bootstrap final weights equal to zero. Now, within each bootstrap sample at least one cluster per stratum will have bootstrap final weights equal to zero, so that members of the same cluster cannot be identified by their zero weights. Unfortunately, when the bootstrap final weights are combined over all B bootstrap samples, cluster membership can be identified. By grouping individuals based on zero and non-zero bootstrap final weights, the members of each cluster can be identified. Although location of the clusters is not given on the PUMF, use of other variables on the PUMF may allow users to deduce the location of a cluster, thus breaching confidentiality.

This problem occurs because under a stratified multi-stage design the bootstrap resamples entire clusters. In the case of stratified simple random sampling, confidentiality is preserved since the cluster consists of a single element. Unfortunately, for stratified multi-stage samples (commonly used in social surveys), the bootstrap method does not meet Statistics Canada's confidentiality guidelines.

As a possible solution to this problem, it was suggested to change the size of the simple random sample, m_h, so that equation (4.2) does not reduce to (4.3). Reducing m_h to be less than $n_h - 1$ may cause problems as it is common to select only 2 clusters per stratum. In this case, some sort of collapsing would be necessary to increase the number of clusters per stratum. Increasing m_h to be greater than $n_h - 1$ would result in negative bootstrap weights, which is not problematic as long as the analysts restrict the use of the negative weights to variance estimation and not for point estimation. Upon closer examination it was noted that the zero weights obtained by using $m_h = n_h - 1$ were replaced by negative weights and the problem with confidentiality still persists.

Two solutions have been investigated to resolve the confidentiality problem: (1) modifying the poststratification adjustment, and (2) replacing the bootstrap weight by an average bootstrap weight. The modified poststratification adjustment is given in Yung (1997). The average bootstrap weight method is described below.

4.1 Mean Bootstrap Weights

The confidentiality problem occurs because m^*_{hi} is always equal to zero for one or more clusters. To avoid this problem, produce R bootstrap samples and average the m^*_{hi}'s over the R samples. As long as each cluster appears in at least one of the R bootstrap samples, the averages will all be non-zero. The steps to perform the mean bootstrap are as follows:

(i) Independently, in each stratum, select a simple random sample of n_h - 1 clusters with replacement from the n_h sample clusters.

(ii) Repeat step (i) R times.

(iii) Let $m^*_{hi(r)}$ be the number of times the *(hi)*-th cluster is selected in the *r*-th bootstrap sample. Let $m^*_{hi(\cdot)} = (1/R)\sum_r m^*_{hi(r)}$ be the average number of times the *(hi)*-th cluster is selected over the *R* bootstrap samples.

(iv) Define the mean bootstrap weight as

$$w^*_{hik(\cdot)} = \frac{n_h}{n_h - 1}\, m^*_{hi(\cdot)}\, w_{hik}.$$

(v) Obtain the bootstrap final weights, $\tilde{w}^*_{hik(.)}$, by performing the same weight adjustment substituting the mean bootstrap weight, $w^*_{hik(\cdot)}$ for the design weight, w_{hik}.

(vi) Calculate θ^* using the bootstrap final weights in the formula for $\hat{\theta}$.

(vii) Independently replicate steps (i) to (vi) a large number of times, *B*, and calculate the corresponding estimates, $\tilde{\theta}^*_{(1)}, \dots, \tilde{\theta}^*_{(B)}$.

The mean bootstrap variance estimator is then given as

$$v_{MB}\left(\hat{\theta}\right) = \frac{R}{B}\sum_b \left(\tilde{\theta}^*_{(b)} - \tilde{\theta}^*_{(\cdot)}\right)^2$$

where $\tilde{\theta}^*_{(\cdot)} = (1/B)\sum_b \tilde{\theta}^*_{(b)}$. We note that the size of *R* should be large enough so that the chance of $m^*_{hi(r)} = 0$ for all $r = 1, \dots, R$ is very small, but it should not be so large that drawing $R \times B$ bootstrap samples becomes computationally unfeasible.

To justify the mean bootstrap variance estimator, we consider the linear case (*i.e.*, $\hat{\theta} = \hat{Y}$ where $\hat{Y}$ is given by equation (4.1)). Letting E_* denote expectation with respect to bootstrap sampling, we wish to evaluate

$$E_*\left(v_{MB}(\hat{\theta})\right) = \frac{R}{B}\sum_b E_*(\tilde{Y}_{(b)} - \tilde{Y}_{(\cdot)})^2$$

where $\tilde{Y}_{(b)} = \sum_{(hik)\in s} w^*_{hik(\cdot)}\, y_{hik}$ and $\tilde{Y}_{(\cdot)} = (1/B)\sum_b \tilde{Y}_{(b)}$. Note that $w_{hik(\cdot)}$ depends on *b*, but for notational simplicity, we drop the subscript *b*. Replacing $\tilde{Y}_{(\cdot)}$ with $\hat{Y}$ (the two are asymptotically equivalent, Rao and Wu 1988), we have,

$$E_*(\tilde{Y}_{(b)} - \hat{Y})^2 = E_*(\tilde{Y}^2_{(b)}) - 2\hat{Y}E_*(\tilde{Y}_{(b)}) + \hat{Y}^2. \qquad (4.4)$$

The $m^*_{hi(r)}$ follow a multinomial distribution with parameters $(n_h - 1)$ and $p_i = (1/n_h)$ for all *i* and *r*. Thus

$$E_*(m^*_{hi(\cdot)}) = E_*(m^*_{hi}) \qquad (4.5)$$

where $m^*_{hi} \sim M((n_h - 1)\,, p_i = 1/n_h,$ for all *i*). Similarly, we obtain the following bootstrap moments:

$$E_*(m^{*2}_{hi(\cdot)}) = \frac{1}{R}V_*(m^*_{hi}) + \left[E_*(m^*_{hi})\right]^2,$$

$$E_*(m^*_{hi(\cdot)}\, m^*_{hj(\cdot)}) = \frac{1}{R}C_*(m^*_{hi}, m^*_{hj}) + \left[E_*(m^*_{hi})\right]^2, \qquad (4.6)$$

$$E_*(m^*_{hi(\cdot)}\, m^*_{gi(\cdot)}) = E_*(m^*_{hi})\, E_*(m^*_{gi}) \text{ and}$$

$$E_*(m^*_{hi(\cdot)}\, m^*_{gj(\cdot)}) = E_*(m^*_{hi})\, E_*(m^*_{gj})$$

where V_* and C_* denote variance and covariance with respect to bootstrap sampling respectively. Substituting expressions (4.5) and (4.6) in equation (4.4) gives, after some simplification,

$$E_*(\tilde{Y}_{(b)} - \hat{Y})^2 = \sum_h \sum_i \left(\frac{n_h}{n_h - 1}\right)^2 \frac{1}{R}V_*(m^*_{hi})\, y^2_{hi} +$$

$$\sum_h \sum_i \sum_{j,j\neq i} \left(\frac{n_h}{n_h - 1}\right)^2 \frac{1}{R}C_*(m^*_{hi}, m^*_{hj})\, y_{hi}y_{hj}.$$

Noting that $V_*(m^*_{hi}) = ((n_h - 1)/n_h)^2$ and $C_*(m^*_{hi}, m^*_{hj}) = -((n_h - 1)/n^2_h)$, we finally have

$$E_*(\tilde{Y}_{(b)} - \hat{Y})^2 = \frac{1}{R}\sum_h \frac{n_h}{n_h - 1}\sum_i (y_{hi} - \frac{1}{n_h}\sum_j y_{hj})^2$$

$$= \frac{1}{R}v(\hat{Y}),$$

where $v(\hat{Y})$ is a commonly used variance estimator (see Yung and Rao 1996). Hence, in the linear case the mean bootstrap variance estimator reduces to the customary variance estimator, $v(\hat{Y})$.

5. EMPIRICAL COMPARISONS

Although consistency of the bootstrap variance estimator has been previously established, the small sample performance of the bootstrap variance estimator was compared to the CV look-up tables empirically. For this comparison, the PUMF of Statistics Canada's 1994 National Population Health Survey (NPHS) was used.

5.1 National Population Health Survey

The NPHS was designed to collect information related to the health of the Canadian population. The objectives of the NPHS included:

- to aid in the development of public policy by providing measures of the level, trend and distribution of the health status of the population;
- to provide data for analytic studies that will assist in understanding the determinants of health;
- to increase the understanding of the relationship between health status and health care utilization, including alternative as well as traditional services.

The design of the NPHS consisted of a stratified two-stage design. In the first stage, homogeneous strata were formed and independent samples of clusters were drawn from each stratum. In the second stage, dwellings were selected within each sampled cluster. The design weights were obtained based on both stages of sampling. To obtain the final weight, a series of twelve weighting adjustments were performed with the last adjustment being a poststratification adjustment. More information on the NPHS design and weighting is available in the NPHS PUMF documentation.

To implement the bootstrap, $n_h - 1$ clusters were sampled with replacement within each stratum. One hundred sets of mean bootstrap weights were generated, with each mean weight based on 20 bootstrap samples (*i.e.*, 2000 bootstrap samples in total). For each set of mean bootstrap weights, only the poststratification adjustment was performed. For comparison purposes, a 'true CV' was computed based on the full jackknife variance estimator (without collapsing), since the jackknife method is currently in use for the NPHS. Table 5.1 gives some results of the comparison between the CV's obtained from the bootstrap, collapsed jackknife and CV look-up tables for means, totals and ratios of categorical and continuous variables.

Table 5.1
Comparison of CV's for Means, Totals and Ratios for the NPHS

$CV_i - CV_J$	Bootstrap	Collapsed Jackknife	CV table
± 1%	64 (86.8%)	41 (45.6%)	45 (50.0%)
± 2%	82 (91.1%)	62 (68.9%)	58 (64.4%)
± 3%	88 (97.8%)	72 (80.0%)	64 (71.1%)
± 4%	90 (100.0%)	77 (85.6%)	71 (78.9%)
> 4%		90 (100.0%)	75 (83.3%)

In Table 5.1, CV_i denotes the CV based on the bootstrap, collapsed jackknife or the look-up tables and CV_J represents the true CV. Table 5.1 shows one of the drawbacks of the CV look-up table method. Of the 90 estimates, 15 involved a continuous variable which could not be handled by the CV look-up tables. Of the remaining 75 estimates, 71 were within ± 4% of the true CV. The bootstrap and the collapsed jackknife both performed well for all estimates with the bootstrap performing better than the collapsed jackknife (87% versus 46% of the estimates were within ± 1% and 100% versus 86% of the estimates were within ± 4%).

To illustrate the versatility of the bootstrap method, CV's were calculated for the regression coefficients for the model relating a person's health status to a measure of the restriction of activities, age, type of drinker and household income. The model was fit separately within five provinces giving a total of 25 parameter estimates. Again, the full jackknife CV was considered to be the true CV. The results of this comparison are given in Table 5.2.

Table 5.2
Comparison of CV's for Regression Coefficients for the NPHS

$CV_i - CV$	Bootstrap	Collapsed Jackknife
± 1%	13 (52.0%)	6 (24.0%)
± 2%	19 (76.0%)	11 (44.0%)
± 3%	20 (80.0%)	15 (60.0%)
± 4%	21 (84.0%)	18 (72.0%)
> 4%	25 (100.0%)	25 (100.0%)

From Table 5.2, we see that the bootstrap performs consistently better than the collapsed jackknife with 52% of the estimated CV's within ± 1% for the bootstrap compared with only 24% for the collapsed jackknife. In addition, the bootstrap has slightly more estimates within ± 4% (84% versus 72%).

6. CONCLUSIONS

The addition of bootstrap final weights to PUMF's will allow users to calculate correct design-based variance estimators (and hence CV's) for categorical and continuous variables as well as for complex statistics such as regression coefficients. Although some technical knowledge is required, it is felt that the calculation of the bootstrap variance estimators is straightforward. Empirical comparisons with the currently used CV look-up method demonstrates the superiority of the bootstrap method both in terms of accuracy and the types of estimators for which it can be applied. While comparisons with the collapsed jackknife indicate only a slightly better performance for the bootstrap CV's, at this time the bootstrap methodology has better theoretical justifications.

REFERENCES

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Mayda, J.E., Mohl, C., and Tambay, J.-L. (1996). Variance estimation and confidentiality: They are related. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 135-141.

National Population Health Survey Public Use Microdata File Documentation, Statistics Canada Publication Number 82F0001XCB.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rao, J.N.K., Wu, C.F.J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 209-217.

Rust, K. (1986). Efficient replicated variance estimation. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Washington, D.C., 81-87.

Valliant, R. (1995). Limitations of balanced half sampling when strata are grouped. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Washington, D.C., 120-125.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. Springer-Verlag New York Inc.

Yung, W., and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

Yung, W. (1997). Variance estimation for public use files under confidentiality constraints. *Proceedings of the Survey Research Methods Section, American Statistical Association*, Washington, D.C., to appear.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A WITHOUT REPLACEMENT RESAMPLING PROCEDURE FOR SURVEY DATA

J.C.S. Cabeça[1]

ABSTRACT

Application of resampling procedures to survey data has been widely studied for the last years, and presents some special problems. A naïve application of Efron's standard bootstrap does not reflect the properties of without replacement sampling from finite populations. Otherwise, without replacement bootstrap methods require $k = N/n$ replications of the sample s in order to create a pseudopopulation of size N. If k is not integer, a randomization is needed, and bootstrap pseudopopulation is somewhat artificial (Rao and Wu 1988). We present an adjustment to Gross' (1980) method, that allows us to apply without replacement bootstrap and requires no randomization. However, the proposed method requires more calculation.

KEY WORDS: Survey sampling; Replication procedures; Bootstrap distribution; Bootstrap moments.

1. INTRODUCTION

Let $U = \{1, ..., i, ..., N\}$ be a finite population consisting of N distinguishable units, divided in L nonoverlapping strata $U_1, U_2, ..., U_L$ of sizes $N_1, N_2, ..., N_L$ respectively, with $\sum_{h=1}^{L} N_h = N$. We denote $y_i, i \in U_h, h = 1, ..., L$, the measurements of a characteristic Y for the i-th unit in the h-th stratum. We assume that the values $y_i, i \in U$, are unknown before sampling is done. We are interested in estimation of a general parameter $\theta = \theta(Y)$, where $Y = \{y_i : i \in U_h, h = 1, ..., L\}$, usually θ being the (unknown) population mean

$$\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h,$$

where $W_h = N_h/N$ and $\bar{Y}_h = \sum_{i \in U_h} y_i/N_h$.

A stratified simple random sample of size n consists in selecting independently in each stratum a simple random sample without replacement s_h of size n_h with $\sum_{h=1}^{L} n_h = n$. A natural estimator of $\theta = \theta(Y)$ is $\hat{\theta} = \hat{\theta}(y)$, where $y = \{y_i : i \in s_h, h = 1, ..., L\}$. When $\theta = \bar{Y}, \hat{\theta}$ may be given by the sample mean

$$\bar{y} = \sum_{h=1}^{L} W_h \bar{y}_h, \qquad (1)$$

where $\bar{y}_h = \sum_{i \in s_h} y_i/n_h$, with variance

$$\mathrm{Var}(\bar{y}) = \sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} S_h^2$$

estimated without bias by

$$\mathrm{var}(\bar{y}) = \sum_{h=1}^{L} W_h^2 \frac{1 - f_h}{n_h} s_h^2, \qquad (2)$$

where $S_h^2 = \sum_{i \in U_h}(y_i - \bar{Y}_h)^2/(N_h - 1), s_h^2 = \sum_{i \in s_h}(y_i - \bar{y}_h)^2/(n_h - 1)$ and $f_h = n_h/N_h$.

Inference on mean and other linear estimators is usually done on basis of normal approximation to the estimator's distribution. However, when the estimator has a more complicated formula, its mean square error (*MSE*) and distribution may be difficult to obtain. In those situations, we can use bootstrap as a device to study properties like variance or bias of the estimators, and construct confidence intervals. To validate bootstrap estimators, we usually analyse the linear case, comparing their performances to the performance of estimators based on normal approximation.

2. EXISTING BOOTSTRAP METHODS

One major difficulty when dealing with complex survey data is that finite population and sampling design (*e.g.*, without replacement sampling, unequal probabilities, multistage sampling) often induce a non-*i.i.d.* structure to the data (Sitter 1992). The bootstrap sampling design should mimic the original scheme as close as possible. Furthermore, the bootstrap variance should match the usual variance estimate in the linear case, in order to establish the good performance and the validation of resampling procedures. These two goals led research in two main directions: I) the bootstrap without replacement (*BWO*) and ii) the bootstrap with replacement (*BWR*).

Gross (1980) first presented a *without-replacement* bootstrap procedure for survey data in the case of a single stratum. The idea was to mirror the original sampling design and to recover the finite population correction $(1 - f)$ in variance formula (Rao and Wu 1988). His method consists in replicating the sample s k times (where $k = N/n$, and k integer), in order to get a bootstrap pseudopopulation U^* of size N, the same size as the original population. The fact that k is seldom an integer is a strong restriction to this method; furthermore, this method does not match the usual unbiased variance estimator in the linear

[1] Júlio César S. Cabeça, Laboratoire de Méthodologie du Traitement des Données, Université Libre de Bruxelles, C.P. 124, Avenue Jeanne 44, 1050 Bruxelles, Belgique; e-mail: jucabeca@ulb.ac.be.

case, but this problem disappears if we consider the bootstrap sample size $n_h - 1$ instead of n (Rao and Wu 1988).

Bickel and Freedman (1984) proposed an extension to Gross' procedure, for situations where the bootstrap sample size in each stratum is not an integer, and allows an extension to $L > 1$. They proposed to construct a pseudo-population in the stratum $h, h = 1, ..., L$, by replicating k_h times (where k_h is the integer part of N/n) the sample s_h with probability $0 < \alpha_h \leq 1$ and $k_h + 1$ times the sample s_h with probability $1 - \alpha_h$, where

$$\alpha_h = \left(1 - \frac{r_h}{n_h}\right)\left(1 - \frac{r_h}{N_h - 1}\right)$$

satisfy the condition

$$\frac{\alpha_h}{k_h} + \frac{1 - \alpha_h}{k_h + 1} \approx f_h,$$

that is the average resampling fraction in the h-th stratum is close to the true value f_h. In the above expression, $r_h = N_h - k_h n_h$. The bootstrap sample s^* consists of L simple random samples of sizes $n_1, ..., n_L$, selected from the pseudopopulations $U_1^*, ..., U_L^*$, independently in each stratum.

Chao and Lo (1985) proposed a similar strategy to Bickel and Freedman, but with a different randomization to handle noninteger resample sizes, in the case where N is not an integral multiple of n. They considered the case of a single stratum. Let $N_1 \leq N \leq N_2$ be the two nearest multiples of n. If we randomize between N_1 and N_2 with probability α and $1 - \alpha$ respectively, both the bootstrap mean and variance match the mean and variance in the original sampling design, where α should be such that

$$F(N) = \alpha F(N_1) + (1 - \alpha) F(N_2)$$

and

$$F(t) = \left(1 - \frac{n}{t}\right) \frac{t(n - 1)}{(t - 1)n}.$$

Booth, Butler and Hall (1994) also proposed a without replacement resampling procedure, which differs from those of Bickel and Freedman (1984) and Chao and Lo (1985) in the way of constructing the bootstrap population. The authors considered a stratified random sampling framework. Let k_h be the integer part of

$$\frac{N_h}{n_h} \quad \text{and} \quad r_h = N_h - n_h k_h.$$

The bootstrap pseudopopulation U_h^* in the stratum $h, h = 1, ..., L$, is formed by combining k_h replicates of s_h with a simple random sample of size r_h selected without replacement from s_h. The bootstrap sample is obtained by selecting a random sample of size n_h without replacement independently in each stratum.

In a different direction, McCarthy and Snowden (1985) proposed a *with-replacement* procedure. This approach is closer to the Efron's standard bootstrap than *without-replacement* bootstrap, but it doesn't parallel the original sampling design. In this approach, one applies the standard bootstrap with a general resample size m (usually $m \neq n$); if m is noninteger, they propose a randomization between backeting integers.

Rao and Wu (1988) proposed a rescaling procedure applied to a nonlinear function of means $\hat{\theta} = g(\bar{y})$. In this method, one considers the bootstrap population being the original sample ($U^* \equiv s$), and selects with replacement bootstrap samples of size $m = (n_h - 2)^2/(n_h - 1)$, rescaling each resampled value appropriately so that the resulting variance estimate matches the usual unbiased variance estimate in the linear case. Rao and Wu's method yields consistent bias and variance estimates and the bootstrap histogram matches the second-order term of the Edgeworth expansion of $\bar{y}$, as $L \to \infty$. However, the fact that this method requires rescaling each resampled value at each bootstrap iteration makes that the calculation of rescaling factors requires summary statistics, that may be cumbersome in complex surveys.

The method proposed by Rao and Wu (1988) can also be applied to stratified simple random sampling without replacement, with

$$m_h = [(1 - f_h)(n_h - 2)^2]/[(1 - 2f_h)^2(n_h - 1)],$$

which is not necessarily integer. In the case of a single stratum, their method applied to simple random sampling without replacement approximates the distribution of a t statistic with a remainder order of $0(n^{-1})$.

Sitter (1992) proposed a *mirror-match* bootstrap that combines both *with* and *without-replacement* sampling components, and achieves both goals: (*i*) it mimics the original sampling design and (*ii*) it matches the usual unbiased estimate of variance. His method entails the following steps:

(1) Resample n_h' $(1 \leq n_h' \leq n_h)$ units without replacement from stratum h to mirror the original sampling scheme (*without-replacement* component);
(2) Repeat step 1 $k_h = [n_h(1 - f_h^*)]/[n_h'(1 - f_h)]$ times independently, replacing the resamples of size n_h' each time, where $f_h^* = n_h'/n_h$ and $n_h^* = k_h n_h'$. This is the (*with-replacement* component);
(3) Repeat steps 1 and 2 independently for each stratum to get the bootstrap sample s^*, and calculate the bootstrap estimate θ^*;
(4) Repeat steps 1 - 3 a large number of times, B to get $\hat{\theta}_1^*, ..., \hat{\theta}_b^*, ..., \hat{\theta}_B^*$;
(5) Estimate $\mathrm{var}(\hat{\theta})$ by $E_*(\hat{\theta}^* - \hat{\theta})^2$.

If k_h in step 2 is noninteger, a randomization is needed, by putting K_h a random variable such that

$$\Pr\left[K_h = k_{h1}\right] = \frac{\left(\frac{1}{k_h} - \frac{1}{k_{h2}}\right)}{\left(\frac{1}{k_{h1}} - \frac{1}{k_{h2}}\right)} = p_h$$

$$\Pr\left[K_h = k_{h2}\right] = 1 - p_h,$$

where $1 \leq k_{h1} \leq k_h \leq k_{h2} \leq n_h$, and k_{h1} and k_{h2} integers (usually, k_{h1} is the greatest integer less than k_h and k_{h2} is the smallest integer less than k_h). The randomization must be done independently for each stratum and repeated at each bootstrap iteration.

3. MOTIVATION FOR A NEW METHOD

In section 2, we argued that a naïve application of Efron's standard bootstrap does not properly generate a bootstrap distribution that reflects the characteristics and properties of without replacement sampling from finite populations. We also discussed some problems arising when applying resampling procedures to survey data. *Without-replacement* bootstrap seems to be more intuitively appealing than *with-replacement* bootstrap, but if N is not a multiple of n, the bootstrap pseudopopulation created by randomization between two created populations is somewhat artificial (Rao and Wu 1988).

In this section, we present an adjustment to Gross' approach that allows us to implement a *without-replacement* resampling procedure and needs no randomization to handle noninteger resample size.

3.1 Some Previous Results

For simplification, let us consider in this section one single stratum.

Let T be an integer, and replicate the population U T times, in order to have an expanded population of size TN, $U_{xp} = \{1, ..., N, ..., 1, ..., N\}$, formed by T replications of U. We denote U_t the t-th replication of U, $t = 1, ..., T$. We can easily see that $F_{xp} = F$, where F and F_{xp} are the distribution functions of the populations U and U_{xp}, respectively. Furthermore, we obtain

$$\bar{Y}_{xp} = \frac{1}{TN} \sum_{i \in U_{xp}} y_i$$

$$= \frac{1}{TN} \sum_{t=1}^{T} \sum_{i \in U_t} y_i$$

$$= \bar{Y}$$

and

$$\sigma^2_{xp} = \frac{1}{TN} \sum_{i \in U_{xp}} (y_i - \bar{Y})^2$$

$$= \sigma^2,$$

proving that both parameters mean and variance do not depend on the number of replications T. However,

$$S^2_{xp} = \frac{TN}{TN - 1} \sigma^2.$$

It is clear that S^2_{xp} goes to σ^2 when $T \to \infty$, even if N remains constant. This means that the actual without replacement sampling approaches the conditions of with replacement sampling.

We select a simple random sample of size n' without replacement from U_{xp}. First-order inclusion probabilities don't change when we draw a simple random sample from U in place of U_{xp}. The sample mean $\bar{y}_{xp}$ is still an unbiased estimator of $\bar{Y}$, and now its variance is given by

$$\mathrm{Var}(\bar{y}_{xp}) = \left(1 - \frac{n'}{TN}\right) \frac{S^2_{xp}}{n'},$$

estimated without bias by

$$\mathrm{var}(\bar{y}_{xp}) = \left(1 - \frac{n'}{TN}\right) \frac{s^2_{xp}}{n'}, \qquad (3)$$

where

$$s^2_{xp} = \sum\nolimits_{i \in s_{xp}} (y_i - \bar{y}_{xp})^2/(n' - 1).$$

$\mathrm{Var}(\bar{y})$ is greater when we draw a simple random sample from U_{xp} than it is when we draw it from U, *i.e.*, if $n' = n$, then $\mathrm{Var}(\bar{y}_{xp}) \geq \mathrm{Var}(\bar{y})$; otherwise, if we want that $\mathrm{Var}(\bar{y}_{xp}) = \mathrm{Var}(\bar{y})$, then we need to have

$$n' = \frac{T(N-1)n}{TN - Tn + n - 1}. \qquad (4)$$

In the special cases where $T = 1$, *i.e.*, $U_{xp} \equiv U$, we have $m = n$, and when $T \to \infty$ we have $n' = [f(N-1)]/(1-f)$, which is the relation between with and without replacement sample sizes. In particular, we note that if $f \to 0$, then $n' \approx n$, and if $f \leq ½$, which is the most frequent case, then $n' \leq N - 1$.

3.2 The Proposed Method

We present now a resampling procedure for the estimation of a general parameter θ, that also achieves both previous stated goals and exiges no randomization between bracketing integers to handle noninteger resample size.

Suppose we start from a population U of size N divided in L strata. We replicate each stratum T_h times, $h = 1, ..., L$, to obtain an expanded population U_{xp} of size $\sum_{h=1}^{L} T_h N_h$, and we draw a stratified simple random sample without replacement of size n', consisting of L independent simple random samples of sizes n_h' from U_{hxp}, $h = 1, ..., L$, with $n' = \sum_{h=1}^{L} n_h'$. The algorithm can be enounced as follows:

(1) Replicate k_h times the sample s_h $h = 1, ..., L$, in order to have an expanded sample, *i.e.*, a bootstrap pseudopopulation of size $k_h n_h'$ in each stratum, $s_{hxp} = U_h^*$, formed by k_h replications of s_h, where $k_h = (T_h N_h / n_h'$, and k_h is an integer. This can be possible by putting $T_h N_h$ equal to a multiple of both N_h and n_h'.

(2) Draw a simple random sample without replacement from the bootstrap pseudopopulation in each stratum to get a bootstrap sample s_h^*, and calculate $\hat{\theta}^*$.

(3) Repeat steps 1 and 2 a large number of times, B, to get $\hat{\theta}_1^*, ..., \hat{\theta}_b^*, ..., \hat{\theta}_B^*$.

(4) Estimate $\mathrm{var}(\hat{\theta})$ with

$$v = E_*(\hat{\theta}^* - E_*\hat{\theta}^*)^2$$

or its Monte Carlo approximation $\tilde{v} = 1/B \sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*)^2$, where $\hat{\theta}^{*}_{(\cdot)} = 1/B \sum_{b=1}^{B} \hat{\theta}^{*}_{b}$. Both $E_{*}\hat{\theta}^{*}$ and $\hat{\theta}^{*}_{(\cdot)}$ can be replaced by $\hat{\theta}$.

When $\theta = \bar{Y}$ and $\hat{\theta} = \bar{y}$, it is clear that

$$E_{*}(\bar{y}^{*}) = \bar{y}$$

$$v = \mathrm{Var}_{*}(\bar{y}^{*}) = \sum_{h=1}^{L}\left(1 - \frac{1}{k_h}\right)\frac{s_h^2}{n'_h}. \tag{5}$$

Using the fact that $k_h = f_h^{-1}$, the bootstrap variance estimator (5) reduces to the usual unbiased variance estimate var($\bar{y}$). Note also that if $T_h \to \infty$, so does k_h faster than T_h.

4. MATCHING THIRD MOMENTS

We have seen that $E_{*}(\bar{y}^{*}) = \bar{y}$ and $\mathrm{Var}_{*}(\bar{y}^{*}) = \mathrm{var}(\bar{y})$, *i.e.*, the first two moments of the bootstrap distribution of $\bar{y}^{*}$ match the usual unbiased estimates of the first two moments of $\bar{y}$. We now consider the third moment $\mu_3(\bar{y})$ that can be obtained by applying the results of Sukhatme *et al.* (1984):

$$\begin{aligned}\mu_3(\bar{y}) &= E(\bar{y} - \bar{Y})^3 \\ &= \sum_{h=1}^{L} W_h^3 \frac{(T_h N_h - n'_h)(T_h N_h - 2n'_h)}{n'^2_h (T_h N_h - 1)(T_h N_h - 2)} \mu_{3h}\end{aligned} \tag{6}$$

where $\mu_{3h} = \sum_{i \in U_{hxp}} (y_i - \bar{Y}_h)^3 / (T_h N_h)$. Note that if $T_h \to \infty$, then

$$\mu_3(\bar{y}) = \sum_{h=1}^{L} W_h^3 \mu_{3h} / n'^2_h, \tag{7}$$

the natural result in with replacement sampling. An unbiased estimate of $\mu_3(\bar{y})$ is given by

$$\hat{\mu}_3(\bar{y}) = \sum_{h=1}^{L} W_h^3 \frac{(1 - f_h)(1 - 2f_h)}{(n'_h - 1)(n'_h - 2)} m_{3h}, \tag{8}$$

where $m_{3h} = \sum_{i \in s_h} (y_i - \bar{y}_h)^3 / n'_h$ is the sample third moment, and $f_h = n'_h / (T_h N_h)$. It is clear that

$$\hat{\mu}_3(\bar{y}) = \sum_{h=1}^{L} W_h^3 \frac{m_{3h}}{(n'_h - 1)(n'_h - 2)}, \tag{9}$$

when $T_h \to \infty$, which is the estimator of $\mu_3(\bar{y})$ under with replacement sampling. The third moment of the bootstrap distribution of $\bar{y}^{*}$ is given by

$$\begin{aligned} E_{*}(\bar{y}^{*} - \bar{y})^3 &= \sum_{h=1}^{L} W_h^3 \frac{(1 - \frac{1}{k_h})(1 - \frac{2}{k_h})}{(n'_h - \frac{1}{k_h})(n'_h - \frac{2}{k_h})} m_{3h} \\ &= \sum_{h=1}^{L} W_h^3 A_h \frac{(1 - f_h)(1 - 2f_h)}{(n'_h - 1)(n'_h - 2)} m_{3h}, \end{aligned} \tag{10}$$

where $A_h = [(n'_h - 1)(n'_h - 2)] / [(n'_h - f_h)(n'_h - 2f_h)]$ is a small correction factor that vanishes when $n'_h \to \infty$. Note also that if $k_h \to \infty$, then $E_{*}(\bar{y}^{*} - \bar{y})^3$ goes to $\sum_{h=1}^{L} W_h^3 m_{3h} / n'^2_h$, the bootstrap third moment under with replacement sampling.

5. DISCUSSION

Application of resampling procedures to survey data is not straightforward. The non-*i.i.d.* structure forced by sampling design impose restrictions to resampling procedure in order to reflect as better as possible the particularities of the original sampling scheme. There is probably no ideal solution. We analysed some algorithms presented in literature (Gross 1980; Bickel and Freedman 1984; Chao and Lo 1985; McCarthy and Snowden 1985; Rao and Wu 1988; Booth, Butler and Hall 1994; Sitter 1992). All of them require a large amount of calculation, but this seems to be a minor problem if we accept that price of calculation tends to zero. The proposed method eliminates the randomization between two artificial populations, it requires only one bootstrap pseudopopulation, and requires no transformation of data. However, the size of pseudopopulation is bigger than for other methods presented in this work. Some other research about its performance is being done.

REFERENCES

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Booth, J.G., Butler, R.W., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89, 1282-1289.

Chao, M.T., and Lo, S.H. (1985). A bootstrap method for finite populations. *Sankhyā*, Seri A, 47, 399-405.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications*. (2nd. Ed.), New York: John Wiley.

Gross, S. (1980). Median estimation in sample surveys. *Proceeding Section Survey Research Methods, American Statistical Association*, 181-184.

McCarthy, P.J., and Snowden, C.B. (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, (Seri 2, no. 95), Public Health Service Publication, Washington, DC: U.S. Government Printing Office, 85-1369.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-245.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S., and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Iowa State University Press, 3rd Ed., Ames, Iowa.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A HIERARCHICAL BAYES ANALYSIS OF CENSUS UNDERCOVERAGE

P. Dick and Y. You[1]

ABSTRACT

In 1991, Statistics Canada decided to adjust the Population Estimates Program to account for net undercoverage in the 1991 Census. One approach to estimating the true provincial populations is to use a Hierarchical Bayes model. The first stage model addresses the uncertainty with the survey results, while the subsequent stages address uncertainty with the underlying differences between provincial undercoverage rates. Demographic methods are used to model the expected differences between the provincial undercoverage rates. In addition, a procedure for detecting differences between various Hierarchical Bayes models are proposed.

KEY WORDS: Hierarchical Bayes; Gibbs sampler; Predicitive distribution.

1. INTRODUCTION

In 1991, in a departure from the established procedure, it was decided to revise the population estimates to agree with the Census counts adjusted to account for net undercoverage – the estimated difference between the number of persons missed and the number of erroneous inclusions in the Census. The new base population was formed by adding the net provincial undercoverage estimate to the provincial Census count. This created an adjusted base upon which all the other population figures were derived using modelling and demographic methods.

The population estimates in 1991 were based on the Census counts adjusted for the estimated net undercoverage in the Census. The technical criteria for adjustment resulted in a procedure known as the preliminary test. This test uses the results of the coverage studies to decide between a full adjustment and no adjustment. The results of this procedure indicated that the Census counts with the net undercoverage added in at the provincial level were an improvement on the Census counts alone with regard to, both, the estimates of the provincial populations and to the estimates of the provincial shares of the national population.

After the release of the 1991 coverage studies' results, a debate was started on examining various estimators of the provincial undercoverage. Rivest (1996) presented a composite estimator that used the national undercoverage rate as a synthetic estimate. The effect of this composite estimator was to shrink all provincial undercoverage rates to the national rate – with each province shrinking by a fixed ratio of the differential provincial undercoverage rate. Thus provinces close to the national rate moved relatively little while provinces far from the national rate moved by a more substantial amount.

In recent years considerable attention has been given to applying calculation intensive methods to improve the direct estimate of net undercoverage from sample surveys. Datta (1992) proposed hierarchical and empirical Bayes methods for the adjustment of the United States census based on the results of the Post Enumeration Survey. Dick and You (1997) examined a Hierarchical Bayes model for estimating the provincial undercoverage rates in the 1991 Canadian Census. This paper is an extension of Dick and You and will examine methods of model checking and validating in an Hierarchical Bayes setting.

2. BAYESIAN MODEL

As in Dick and You, we suppose there are n provinces and in the i-th province the Census has counted Y_i persons while an unknown number U_i persons were missed by the Census. The coverage studies provide an estimate, $\hat{U}_i$, of the net undercoverage along with an associated (known) variance, ξ_i^2. One objective of the Population Estimates Program is to estimate the true population of the i-th province on Census Day.

The true population of the i-th province is written as $T_i = Y_i + U_i$. Since the Census count is observed without sampling error most of the work in constructing a model centres around the estimate of missed persons. The coverage studies use a sample survey which through standard estimation procedures produce an estimate of the missed persons. The model generally used to describe this estimation situation is written as

$$\hat{U}_i = U_i + \varepsilon_i,\ i = 1, ..., n, \tag{2.1}$$

where $\varepsilon_k \sim N(0, \xi_i^2)$. This model assumes that the estimators $\hat{U}_i$ are design unbiased, normally distributed with known sampling variances. These assumptions may be restrictive – in particular the estimates of missed persons, $\hat{U}_i$, are certainly subject to (unknown) bias. The assumed

[1] Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario,Canada. K1A 0T6; and Yong You, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada.

distribution (Normal) of the sampling errors, at the province level, seems quite reasonable because of the Central Limit Theorem.

Interpreting the differences in missed persons between provinces is difficult because of the large differences in provincial sizes, so a more meaningful measure is the undercoverage rate which is defined as $r_i = U_i / (U_i + Y_i) = U_i / T_i$. This transformation implies that survey model can be written as

$$\hat{r}_i = r_i + e_i, \; i = 1, ..., n \tag{2.2a}$$

where we take $\hat{r}_i = \hat{U}_i / \hat{T}_i$ and we assume the sampling errors are $e_i \sim N(0, \psi_i^2)$. The sampling variance are related to the original variance ζ_i^2 by $\psi_i^2 = \xi_i^2 (1 - r_i\ )^2 / \hat{U}_i^2$ and are assumed known.

Suppose the true undercoverage rates are related to a number, say p, variables $x_i = (x_{i1}, ..., x_{ip})'$ specific to the i-th province (Fay and Herriot 1979). In particular assume that the linear relationship

$$r_i = \chi_i \beta + v_i, \; i = 1, ..., n \tag{2.2b}$$

is true. Here β is the vector of regression coefficients and v_i are independent and identically distributed random variables with $E(v_i) = 0$ and $V(v_i) = \sigma_v^2$.

Combining these two equations (2.2), we obtain the general mixed linear model

$$r_i = \chi_i \beta + v_i + e_i, \; i = 1, ..., n. \tag{2.3}$$

Note both design induced random variables, e_i, and model based random variables, v_i, are included in the model. An extensive discussion on this model can be found in Ghosh and Rao (1994). As in Ghosh and Rao, we assume that the sampling variance, ψ_i^2, is known allowing for unknown sampling variance is briefly discussed in the conclusions to this paper.

The fixed effect part of the general model relates the true undercoverage rate to underlying set of auxiliary variables. The simplest model sets the regression component equal to some fixed value. A logical choice would be the national undercoverage rate, $\bar{r} = \sum U_i / \sum T_i$, hence we would equate the regression component to the national coverage rate, $x_i' \beta = \bar{r}$. Using the fact that $r_i = (\bar{r} - \sum \alpha_k p_k) + \alpha_i$ where α_i is pre-specified, Dick and You show the general model (2.3) can be written as

$$\hat{r}_i = \bar{r} - \sum \alpha_k p_k + \alpha_i + v_i, + e_i, \; i = 1, 2, ..., n. \tag{2.4}$$

The advantage of this approach is that, for any two provinces, the undercoverage rates will differ by

$$r_i - \ r_j = \alpha_i - \alpha_j. \tag{2.5}$$

This implies that the vector α is just the difference in undercoverage between the i-th province and a fixed value. Thus by pre-specifying how the Census undercoverage is expected to differ between provinces we can specify the fixed component without regard for national undercoverage rate, $\bar{r}$.

Having determined the form of the fixed effects part of the model, the random effects will now be introduced. The basic framework of the Hierarchical Bayes model uses the concept of exchangeability. The first stage, assumes that the undercoverage rates for each province are unbiasedly estimated with known sampling variance. The second stage assumes that the true undercoverage rates for each province are unbiasedly estimated by the national undercoverage rate adjusted for differences between the provinces with an unknown variance. The errors associated with this stage are assumed to be exchangeable: that is the prior opinion on the unexplained portion of the undercoverage rate for Ontario would be the same as for Prince Edward Island. The final stage assumes that the national undercoverage rate has a known distribution.

More formally, this model can be written as:

(i) *The Sampling Model:*

$$[\hat{r}_i \mid r_i] \sim N(r_i, \psi_i^2) \text{ for } i = 1, ..., n \text{ and } \psi_i^2$$

are the known sampling variances;

(ii) *The Population Model*:

$$[r_i \mid \bar{r}, c_i, \sigma_v^2] \sim t(\bar{r} + c_i, \sigma_v^2, \eta_i) \text{ where } c_i = \alpha_i - \sum \alpha_k p_k$$

is the fixed effect part discussed in Section 2.2, σ_v^2 is the population variance and η_i are the (known) degrees of freedom associated with the t-distribution;

(iii) Prior distributions:

$$\bar{r} \sim N(r_o, \Psi_o) \text{ where } r_o \text{ and } \Psi_o \text{ are known and}$$

$$\sigma_v^2 \sim \Gamma(\frac{1}{2} a, \frac{1}{2} b) \text{ where } a \text{ and } b \text{ are known.}$$

The population model can be represented in a form more useful for implementing the Gibbs sampler. It can be shown that the assumed t-distribution can be written in two stages as, first, as normal distribution and then, using a new variable defined in the normal distribution as a Gamma distribution. Note, the degrees of freedom for the t-distribution, η_i, are assumed known. Writing the population in these two parts implies that the full model requires five conditional distributions to completely describe it. The details of this can be found in Dick and You.

The Gibbs sampler is a technique for extracting the marginal distributions from the full conditional distributions instead of the full distribution. Instead of calculating the marginal distribution directly, the Gibbs sampler simulates the results of drawing from the appropriate target distribution. Hence, the posterior expectation and posterior variance can be calculated by, first, simulating a large number of draws from the distribution, and, secondly, calculating the expected value and variance of these draws from the target distribution. A straight forward explanation of the Gibbs Sampler can be found in Casella and George (1992).

The Gibbs sampler needs the full conditionals (see Dick and You). The Gibbs sampler used in the analysis of the model described above was the Bayesian inference Under Bayes Sampling (Spiegelhalter *et al.* 1996). The algorithm is relatively simple:

(1) Draw the true undercoverage rate, $r_i^{(1)}$, using starting values $\zeta_i^{(0)}$ and $\bar{r}^{(0)}$ from the full conditional distribution;

(2) Draw the national undercoverage rate, $\bar{r}^{(1)}$, using starting values $\zeta_i^{(0)}$ and $r_i^{(1)}$ from the full conditional distribution;

(3) Draw the variable $\zeta_i^{(1)}$ using starting values $\sigma_v^{2(0)}$ and $\bar{r}^{(1)}$ from the full conditional distribution;

(4) Finally, draw the variable $\sigma_v^{2(1)}$ using starting values $\zeta_i^{(1)}$ from the full conditional distribution.

Running all four parts completes one cycle of the algorithm, the posterior expectation and posterior variance shown in the next section are the results of completing 12,000 cycles. In order to ensure that the distribution in which the inferences are made is the correct one, a "burn-in" of the first 2,000 cycles was discarded: in effect only the last 10,000 simulations are kept for the analysis. Details can be found in Dick and You.

3. CENSUS UNDERCOVERAGE ESTIMATION

Since 1966, the Reverse Record Check has been the survey vehicle used by Statistics Canada to measure gross number of persons missed by the Census. Starting in 1991, an Overcoverage Study, was conducted to measure the gross number of persons erroneously included in the Census. Together these coverage surveys provided an estimate of the net number of persons missed by the Census. Through the analysis of the results of these surveys, the collection methodology is adjusted in order to improve coverage in the succeeding Census. Details on the coverage studies can be found in Germain and Julien (1993).

The survey had a sample of 56,000 in 1991 and was designed to estimate the number of missed persons in the Census. The sample was allocated to each province to ensure that the maximum standard error on the undercoverage rate would be less than 0.35%; the rest of the sample was allocated proportionally to population. The design was a stratified random sample with a disproportionate sample amongst young adults (20 to 29) – a group more prone to be missed. The sample allocation should be sufficient to give reliable estimates for the provinces and for national age and sex totals.

The model used is described in Section 2 (see Dick and You for full details). However, the values for the fixed effects part of the model α_i and the degrees of freedom η_i in the population model need to be specified. Four different approaches were taken when pre-specifying the fixed effects. First, assume that all provinces have exchangeable errors when sampled from the national undercoverage rate. This model (denoted Model 1) implies that all $\alpha_i = 0$. Secondly, assume that Ontario has an undercoverage rate that is 1% higher and Prince Edward Island has an undercoverage rate that is 1.5% lower than the other 8 provinces (Model 2). This approach is accounting for the fact that it is becoming relatively difficult to conduct a Census in Ontario, and in particular in Toronto because of its size, diversity and complexity. On the other hand, this approach allows for the apparent ease in which a Census can be conducted in Prince Edward Island due to small size and a stable and homogeneous population. Thirdly, assume that the gross undercoverage rates in 1986 are valid estimates of α (Model 3). Finally, assume the population estimates projections determine α (Model 4). Table 1 shows the values for the model specifications.

The population model also needs to have the degrees of freedom (η_i) specified. Three separate assumptions were made concerning the degrees of freedom. First, in order to provide a benchmark series, we have assumed a "very large – $\eta_i = 200$" number for the degrees of freedom for each province; this in effect assumes the population model has a normal distribution. Then, following the suggestion of Datta and Lahiri, we have assigned one degree of freedom (a Cauchy distribution) to the minimum and maximum observations – corresponding to PEI and Ontario: for the other provinces we have assumed 5 and 15 degrees of freedom. For the prior distributions we follow closely the suggestion by Hobert and Casella (1996). They recommend avoiding improper posteriors by using proper priors. They state that "ignorance can be modelled by using a normal prior with large variance and inverted gamma priors with small parameter values for the variance components". Hence, for the national undercoverage rate, we assumed that $\bar{r} \sim N(0.02865, 1)$ which is the observed rate in 1991 with a large variance and for the population variance, we assumed $\sigma_v^2 \sim \Gamma(0.0001, 0.0001)$ which under our model assumes that this quantity is distributed as an Inverse Gamma with small parameter values.

The results of the Gibbs Sampler for the various models are displayed in Table 1: since there is little difference in the estimates for the different degrees of freedom, only the results for $\eta_i = 15$ are presented; the complete results are reported in You (1997). Model 1 shows a general shrinking toward the national rate of 2.86% for all provinces. Provinces close to the national rate such as Quebec and British Columbia move very little while those "far" from the national rate, such as Ontario and PEI move more. Model 2 shows very little movement for any province – essentially showing that the fixed effects part is successful in explaining the variation. Model 3 shows a very large increase in the British Columbia estimate due to the large α value from the 1986 gross undercoverage study. Model 4 shows results very similar to Model 1 – reflecting the difficulty the population estimates projections have detecting inter-provincial migration.

4. MODEL CHECKING AND VALIDATION

One way to assess the adequacy of the above models is through predictive distributions. This is defined by supposing that we assume a prior distribution of $\theta \sim \pi(\theta)$ and we have $x \sim f(x|\theta)$ and $z \sim g(z|\theta)$ then the predictive distribution of z after the observation x (Robert 1994, page 143) is

Table 1
Observed, Prior and Posterior Estimates of Undercoverage

Prov	Obs	Model 1		Model 2		Model 3		Model 4	
		a_1	HB	a_2	HB	a_3	HB	a_4	HB
Nfld	1.99	0	2.08	0	2.08	-1.3	1.92	0.2	2.12
PEI	0.93	0	1.00	-1.5	0.94	-1.1	1.02	0.1	1.01
NS	1.89	0	2.02	0	2.02	-1.1	1.86	-0.6	1.94
NB	3.25	0	3.15	0	3.02	-0.5	3.04	0.1	3.15
Que	2.61	0	2.61	0	2.59	-0.3	2.60	-0.4	2.59
Ont	3.64	0	3.55	1.0	3.60	0.2	3.56	0.5	3.59
Man	1.86	0	2.00	0	2.00	-0.3	1.99	-0.3	1.98
Sask	1.80	0	1.92	0	1.93	-0.8	1.83	0.0	1.95
Alta	2.00	0	2.07	0	2.07	-0.2	2.07	-0.5	2.04
BC	2.73	0	2.74	0	2.69	1.3	2.84	-0.2	2.73

$$g(z|x) = \int g(z|\theta)\,\pi(\theta|x)\,d\theta. \qquad (4.1)$$

This is the conditional density of the observable averaged against the prior knowledge about the "unobservables" (Gelfand 1996). This implies that a useful measure would be the likely value of an observation when the model is fitted to all the observations except the one being examined.

Suppose $\hat{r}_{(i)}$ is the vector of net undercoverage rates for all provinces except the i-th, then using (4.1), the cross-validation predictive density $f(\hat{r}_i \mid \hat{r}_{(i)})$ suggests likely values for the i-th province when the model is fitted to all the observations except the i-th. Gelfand (1995) gives an extended discussion of this measure. One measure that uses this approach, is when we take the expected value of the predictive density and combine it with the observed value, then a 'residual' can be formed from

$$d_i = \hat{r}_i - E(\hat{r}_i \mid \hat{r}_{(i)}). \qquad (4.2)$$

More details on the calculation of this quantity using the results of the Gibbs Sampler can be found in You (1997).

You presents the deviations resulting from the 4 different models when it is assumed that the population component has 5 degrees of freedom. Clearly, there is very little room to choice between the models for Quebec. However, Model 2 really only preform poorly in New Brunswick and Model 3 only poorly in British Columbia. Model 1 and Model 4 are very similar and preform somewhat badly in Newfoundland, PEI, Nova Scotia and Saskatchewan.

Another measure described by Gelfand is the conditional predictive ordinate. This is defined as $c_i = f(\hat{r}_i \mid \hat{r}_{(i)})$ which, when the log is taken is written as

$$\text{cpo}_i = \log f(\hat{r}_i) - \log f(\hat{r}_{(i)}).$$

Hence the cpo contrasts the prior predictive density for all observations with all but the i-th observation. The interpretation that Gelfand (1995, page 153) put on this is "the incremental contribution to the adequacy of the model attributable to the i-th observation". Essentially, we are dealing with the observed likelihood thus a large value of cpo suggests $\hat{r}_i$ a more likely model and a smaller value suggests that does not support the model. Details on using the results of the Gibbs Sampler for calculating the cpo can be found in You (1997).

You also presents the cpo plotted for the four different models for each province. The cpo values are small for New Brunswick indicating none of the models are really adequate. Model 2 has the highest cpo value in every province except Nova Scotia (Model 3), New Brunswick (Model 4) and Saskatchewan (Model 3). No where does Model 1 provide the highest cpo.

5. CONCLUSION AND FUTURE RESEARCH

The Hierarchical Bayes model has shown that some improvement over the direct estimators is possible. The population model is flexible and permits at easy interpretation for the expected differences in provinces. Four different models are considered by specifying the prior knowledge on the special effects. Hierarchical Bayes estimates are provided by Gibbs sampling methods.

Bayesian model checking and choice are illustrated using the cross-validation predictive densities. From residual analysis and cpo comparisons, Model 2 is the most likely model and receives the most support from the data.

Further work needs to be done on the fixed effects in the population model. The impact of the fixed effects component is quite marked but no adequate model has been found yet. Investigations should concentrate on the quality information available from the Census such as non-response rates. Finally, the posterior variance has been underestimated due to the assumption of known sampling errors.

REFERENCES

Casella, G., and George, E.I. (1992). Explaining the Gibbs Sampler. *American Statistician,* 46, 167-174.

Datta, G.S., and Lahiri, P. (1994). Robust hierarchical Bayes estimation of small characteristics in presence of covariates. *Private communication to J.N.K. Rao.*

Dick, J.P., and You, Y. (1997). Bayes and Census Undercoverage. Invited paper Statistical Society of Canada Annual Conference: Fredericton, New Brunswick.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelfand, A.E. (1995). Model determination using sampling based methods. In *Markov Chain Monto Carlo in Practice*, Eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, 145-161. London: Chapman & Hall.

Germain, M.F., and Julien, C. (1993). Results of the 1991 Census coverage error measurement program. *Proceedings of Seventh Annual Research Conference.* United States Bureau of the Census. 55-70.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1995). Introducing Markov chain Monto Carlo. In *Markov Chain Monto Carlo in Practice*, Eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, 1-19. London: Chapman & Hall.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science* 9: 55-93.

Hobert, J.P., and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 91: 1461-1473.

Rivest, L.P. (1996). Some shrinkage estimators for Census undercoverage. Technical Report 96-01, Université Laval.

Robert, C. (1994). *The Bayesian Choice: A Decision-Theoretic Motiviation.* Springer-Verlay, New York.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *BUGS 0.5, Bayesian inference using Gibbs sampling manual.* MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, England CB2 2SR.

You, Y. (1997). *Hierarchical Bayes Models for Census Undercoverage Estimation.* Internal Statistics Canada report.

SESSION I-6

Data Warehousing and Internet Dissemination

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

THE REDESIGN OF THE CANSIM DATABASE

P. Basset and A. Stoyka[1]

ABSTRACT

In this brave new world of Data Warehousing and On-Line Analytical Processing (OLAP) where terms like "multi-dimensionality", "slice and dice" and "hypercube" are on the tip of many database designers' tongues, Statistics Canada's CANSIM II database initiative is exploring new ground not yet charted.

CANSIM II's mandate is to create a repository containing Statistics Canada's aggregate published data together with its corresponding metadata and meta-information. The ability to then effectively navigate this repository is crucial. It is the sheer potential scope of the data content which sets CANSIM II apart from other data warehousing initiatives.

The universes in which most data warehouses operate are actually quite contained. Typically, twenty or thirty individual dimensions for a data warehouse would be considered extensive and once the dimensions have been analyzed and defined, additions are few and far between, and usually quite painful. With CANSIM II, the anticipated number of dimensions is extensive, truly limited to infinity. Additional dimensions are a frequent fact of life. Yet CANSIM II strives to provide a harmonization and organization which not only allows easy navigation, but a view of Statistics Canada aggregate data limited only by the imagination of the client.

The possibility for client's being able to easily bring together data from a number of different sources within Statistics Canada will raise many issues. Past practices of labeling tabulations in isolation will give rise to data appearing to be the same but in fact being different. On the other hand data that is for the same variable may be have a different name and then appear to be different. CANSIM II will be tackling these and other harmonization problems.

This paper will explore the problems encountered by the CANSIM II project from both a technical, "systems design" point of view and from an overall corporate rethinking of how data are defined and connected to meta-data. It will also provide current working solutions evaluated with both their advantages and disadvantages.

KEY WORDS: Relational database; Meta-data; Dissemination.

1. BACKGROUND

On September 14, 1995, the Dissemination Committee of Statistics Canada endorsed a proposal to initiate a project to overhaul the CANSIM-I database. This proposal was recommended by the consultant Ulla De Stricker, whose services had been obtained originally to recommend steps on how to increase the sales of the CANSIM-I CD-ROM. Her findings and judgement were that the sales of CANSIM-I would not increase significantly without significant improvements to the content, structure, scope, and documentation (Meta data) of the CANSIM-I database. By implication, the same holds true for the marketing of the CANSIM-I online distribution through StatCan Online, Internet and secondary distributors. CANSIM-I is a well-established database, but its now 25 year old structure and documentation is outdated and cause for frustration by existing and potential new clients.

As a result, Statistics Canada initiated a multi-year project to redevelop CANSIM and to make CANSIM II the core of a broad data warehouse of all publishable statistics to be used for external data dissemination as well as by subject matter analysts within Statistics Canada. The project team consists of resources from System Development Division and Dissemination Division. The steering committee is composed of Philip Smith – Director General for the PIPES project, Mike Sheridan – Director General for Labour and Household Surveys, Mel Turner – Director of System Development Division and Martin Podehl – Director of Dissemination Division. Work on the design of the database started in April 1996. The database designers considered the existing CANSIM Time Series and the Cross-Classified databases. Analysts and data suppliers within Statistics Canada were consulted as well as those close to external users, including Advisory Services and Marketing. A focus group of major CANSIM users provided additional ideas.

2. MAJOR OBJECTIVES OF THE PROJECT

2.1 Versatile and Robust Database Structure

Create a database that will contain all of Statistics Canada's aggregate data together with its supporting labels. Much of Statistics Canada's aggregate data is on the existing CANSIM Time Series base or the Cross-Classified base; it will be copied over to the new database but into a new structure. The new structure will support time series, hierarchical and cross-sectional data. It is planned that tabulated data which is not at present on either the Cross Classified or the Times Series bases will also be stored on the new CANSIM base.

[1] Penny Basset, Project Manager, Dissemination Division and Ann Stoyka, Database Designer, Systems Development Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

2.2 Full Labeling for Better Search Capabilities

Improve the labeling of tabulated data to make it easier to find and present using more widely recognized and consistent terminology. The existing CANSIM Time Series database structure is restrictive in the field length of its labels. As a result there are many different abbreviations for words which makes searching difficult. The restriction on the length of descriptions required for the two dimensional time series base also contributes to difficulties in finding appropriate data. The proposed multi-dimensional structure will allow for simpler and more consistent labeling.

2.3 Harmonized Labeling of Data

Ease of pulling together data for the same variable from different survey sources will be facilitated by the harmonization of data labels. CANSIM users are often not aware of the different survey sources used to produce data, nor should they need to be. They would like the organization of Statistics Canada to be transparent to them and to be able to easily bring data together into one analytical data table data from different sources. Standard classifications for industries, commodities, occupations, diseases, education and crimes will be available to label data on CANSIM II. Subject matter data providers will be encouraged to label their data using these standard classifications. In addition, standard labels for other variables used by multiple surveys will be available. Examples of such variables include marital status, age groups, gender and family type.

2.4 Intranet and Internet Accessible

Provide access to the database and to Statistics Canada meta-information through a World Wide Web (WWW) interface as well as the existing system of secondary distributors. As the new database becomes populated with data it will be made accessible to Statistics Canada users via the Intranet. This will also provide an opportunity to test search and retrieval functionality before making the database accessible on Statistics Canada's Internet web site.

2.5 Access to Meta-Information

Meta-information about all of Statistics Canada programmes is available on the Statistical Surveys database that is maintained by Standards division. A major project is underway to improve this base of information. The data on CANSIM II will be linked to the meta-information database. In addition, there are the facilities to provide footnoted text down to the data cell level.

3. FEATURES OF THE NEW CANSIM

3.1 Multidimensionality Replaces two Dimensional Time Series

While CANSIM was originally designed to store economic time series data, the new design will house not only time series data but also cross-sectional aggregate data. The new CANSIM database, instead of being of two dimensions, *i.e.*, time and all other variables, will be multi-dimensional. This will allow series to be identified by their separate variable components. For example the series for the CPI for Canada and the provinces with 1981=100 will have three separate dimensions *i.e.*, time, geography and the CPI classification system of goods and services. Each dimension will have a separate name, in this example they would be monthly, Canada and the provinces and the CPI goods and services classification. All data will have geographic and time dimensions. The third type of dimension is referred to as the thematic dimension. Thematic dimensions include standard classifications such as the Standard Industrial Classification, harmonized dimensions such as gender and family type, and dimensions which only apply to one survey, such as employment type which only applies to the labour force survey.

3.2 Consistent and Full Labeling will Allow Better Search Capabilities

Improvement to the labeling of data series is a number one priority for the new CANSIM. Allowing words to be spelt out in full and in phrases will ease keyword searching and allow users to better understand what data they can access. Abbreviations are used inconsistently in the existing CANSIM – full words will eliminate this problem from the start.

Subject matter data suppliers will be encouraged to use harmonized labels. The use of pull down menus to set up the labels for dimensions, their categories and other fields of information will make labeling of data faster than the manual approach of typing in data labels. Harmonization of data labels is not an easy project as each survey is designed for a different purpose. However, some structures can be put in place to avoid obvious inconsistencies. For example, more than one name may exist for the same commodity. The standard commodity classification structure will be readily available on the system for use by author divisions, and the standard name will be the one used in CANSIM.

The categories of the dimensions will appear in their hierarchical relationship, allowing data providers and data users to see the components that make up a data series. While a hierarchical arrangement is available on the present CANSIM time series system, changes, such as new series, are not always incorporated into the hierarchical structure.

3.3 The Aggregate Database for the Statistics Canada Warehouse

A data warehouse is the linking of different computer information databases so that users are able to access many sources of information simultaneously and in an informative way. The new CANSIM will be the macro data area of the Statistics Canada warehouse. This will allow hyperlinks to be made to other information bases in the warehouse, reducing the amount of meta-information that will actually need to be stored in the CANSIM database. This means one-stop shopping for users of statistics Canada's information.

3.4 Single Database to Feed Many Output Formats From Different Data Sources

At present, author divisions produce their tabulations in different computer formats depending upon the dissemination vehicle. For example, on the day before the CPI is released, the tables for the Daily are produced in HTML format, the fax release is in EXCEL format, the tables for publication are produced in PAGEMAKER and another system produces the tables for the publications. This means that if there are last minute changes, then they need to be made in each output format. The CANSIM database is Open Database Connectivity compliant (ODBC); many different output formats will be available, including EXCEL, HTML, dif, Beyond 20/20.

With the new CANSIM, interfaces will be created to allow the DAILY, publication, faxes, CD-ROMs, diskettes, and on-line access all to be available from the same tables within the new CANSIM. When a change is made, it will only need to be made in the source table in CANSIM.

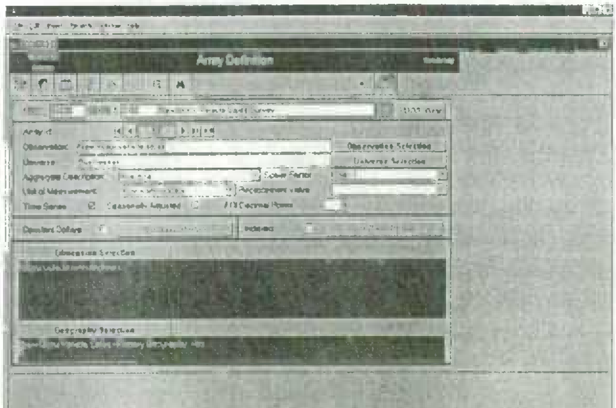
Analytical and subject matter divisions will be able to create new dissemination products from different sources but along a common theme. For example, the SABLE and BID bases will be able to be updated easily from the CANSIM base rather than receiving updates from individual divisions. In the case of BID, twelve divisions are involved. Definitions and methodology references will also be accessible within the same environment as CANSIM, thus allowing meta-information to be easily included in Statistics Canada's products. For example, the definitions of industries, as available from the North American Industry Classification System (NAICS), will be only a mouse click away.

4. ORGANIZATION OF THE NEW DATABASE

The new CANSIM database will store hierarchical multi-dimensional aggregate data (such as that produced by the National Accounts), data tabulated from surveys, and administrative data. It has been built on a relational database platform.

4.1 Array and Observations

Screen #1. Array Definition for New Motor Vehicle Sales by manufacturer and province.



The basic building block of the database is the aggregate statistical estimate that is stored as a data cell. Data cells are grouped into multi-dimensional arrays. Each array consists of data from one survey or statistical programme source, *e.g.*, the balance of payments, the labour force survey. The other condition for data cells to be grouped into the same array is that the statistical estimates contained in one array are the same. This includes the estimate's unit of measurement, scalar factor, number of decimal points and the method of aggregation. See Screen 1 to see the labels required for each array. The statistical estimate of an array is called the observation. Examples of observations include the number of people employed, the value of petroleum products, and the raw material price index.

4.2 Vectors and Databank Numbers

Data cells for the same variables, including geography but changing over time, make up vectors in an array. These vectors are similar to the data series in the CANSIM time series base and are identified by one databank number. They serve as the link to copying data from CANSIM I to CANSIM II. They are also the means by which secondary distributors identify their data series.

4.3 Classifications and Dimensions

4.3.1 Time Dimension

Each array will have at least two dimensions. These are the time and the geography dimensions. The time dimension will be automatically identified by the system as data with their time tags are loaded onto the database. Examples of the time dimension are weekly, monthly, quarterly, annually and occasional.

4.3.2 Geographic Dimension

The geographic dimension is also mandatory for each array. There may be a need for a second geographic dimension. For example the manufacturing origin of the sale of motor vehicles by province has two geographic dimensions. The primary dimension is the location of the sales and the secondary geographic dimension is the location of manufacturing origin. There will be standardized lists of geographic areas available as labels to the data, from which data supplier, can select as appropriate.

4.3.3 Thematic Dimensions, Levels and Members

The thematic dimension is not mandatory but is present for most arrays. An example of an array that does not have a thematic dimension is the annual estimate of the population by province. However, the array for annual population by province and gender has gender as its thematic dimension.

Each dimension has a set of components or attributes that in CANSIM II are called members. A dimension is a collective name for a set of attributes of a variable. For example the dimension name for "men" and "women" is "gender". Conversely the members of the dimension "gender" are "men" and "women". For standard classifications such as the Standard Industrial Classification, the dimension is the SIC, but there also needs to be a version date, such as SIC 1980. The members of the dimension SIC are the names of all the industries contained in the SIC.

There are three kinds of thematic dimensions, namely standard classifications, harmonized dimensions and dimensions that are unique to one survey or statistical programme. The harmonized dimensions tend to be mostly for social concepts such as marital status, gender and family type. The members of the dimension for marital status are: legally married, separated, divorced, widowed, never married, living common-law, not-living common law, living with spouse, not living with spouse. Not that some of these members over-lap but this a accounted for in the hierarchical arrangement of the members. There is certain information that is only collected by one survey or statistical programme. This unique information falls into the non- harmonized group of dimensions. An example of this kind of dimension is motor vehicle manufacturers whose members include North American manufacturers, Japanese vehicle manufacturers and other country vehicle manufacturers.

4.4 Thematic or Topic Organization of Dimensions

To facilitate searching by dimensions when arrays are defined, the dimensions are organized into a thematic hierarchy. The thematic hierarchy on the social side is based upon the social thematic search tool. CANSIM has designed its own thematic organization for the business and economic dimensions.

4.5 Links to Meta-Information Base

The Statistical Data Documentation System (SDDS) number will provide the link to the meta-information databases, where users will be able to access information on the data source. In addition, subject matter data suppliers will be able to provide footnotes as needed at the data cell or the array level.

4.6 Grouping of Arrays to Matrices

Subject matter data suppliers can group arrays with a common theme into matrices. Each matrix has its own title. For example: the arrays for the unemployment rate, employment to population ratio, number in the labour force, employment and unemployed could be grouped into a matrix called labour force indicators. This facilitator serves two functions. Firstly, it saves users from searching for individual arrays to make up a data set for analysis; the searching and assembling work will be done for them in advance. Secondly, the title for the matrix serves as an additional method of identifying relevant data.

5. APPROACH TO DEVELOPING THE DATABASE

5.1 The Prototype 1996/97 and 1997/98

A prototype has been built as a tool to facilitate communications with potential data suppliers and users in order to get feedback and also to put the theories to work. It serves to illustrate major concepts and, like the rest of the system design cycle at this time, is constantly evolving. The initial software is MS Access both for the interactive screens and as the DBMS.

5.2 Testing 1997/98

Six divisions representing different data participated in a CANSIM II test in October 1997. The data sets were: the Raw Material Price Index and Industrial Product Price Index from Prices Division, petroleum products from Manufacturing, Construction and Energy Division, new motor vehicle sales from distributive trades divisions, vital statistics from Health division, the labour force survey from Household Surveys and the Balance of Payments. The test will provide feedback on the design of the database, the screens to set up the arrays, the training, and the copying of data from CANSIM I to CANSIM II. Refinements will be made to the database, the interface to set up the arrays, and the update systems during the last quarter of 1997/98.

5.3 Populating the Base 1998/99 and Web Search and Retrieval Facility

1998/99 will be the most important year for setting up the new arrays of data. This operation will require considerable resources to research and assemble the dimensions to be used by subject matter areas. Each subject matter area will need to identify appropriate people to be trained in the new database structure. The organization of data will need to be reorganized to be stored on CANSIM II. The data will be copied from CANSIM I where possible. CANSIM II will also serve as an archive base. Decisions will need to be taken as to which data should be stored for archival purposes and which for current usage.

SQL Server will be the DBMS for volume testing and possibly some benchmarking. The Production DBMS and Production Hardware has yet to be benchmarked and selected.

A Web Internet search and retrieval facility will be constructed and ,as the database is populated, it will be available for internal Intranet use. The Intranet access tool will serve as internal testing by Statistics Canada's users.

5.4 Security of Access

A number of divisions have asked if they can use CANSIM as their analytical database but this would require secure access to certain data, *e.g.*, data stored prior to its official release or data that is statistically unreliable. A security access system will be designed and built in 19989/99.

5.5 User Information Reporting System

Who is using what data is important from both market research and billing viewpoints. The kinds of information required for each are not always compatible. The design of this system to collect appropriate information and to produce relevant reports will be done in 1998/99.

5.6 Tool to Check Data Integrity

While tools are available on CANSIM I to check data they are not always used. This is for a number of reasons. A research project is underway to identify a suitable method for identifying outliers. Subject matter areas will be asked to check these outliers and to provide explanations as appropriate. The explanations would be attached as footnotes to the data cell. The footnotes will be additional information to help users. For this procedure to be workable it will be important that the number of outliers identified is reasonable.

5.7 Concordance

One of the major frustrations of CANSIM users is the discontinuity of series when classifications are revised. While revisions are inevitable with change in economic structure, information should be provided on the changes to the classification structures. A system has been developed for NAICs which displays the concordances to the previous industrial classifications. It is proposed that the same system be used to store most classification concordances if they are available.

6. EXTERNAL AVAILABILITY OF CANSIM II 2000/2001

The plan is to have the new database fully functional on the Statistics Canada web site by the year 2000/2001. Prior to this, there will be a parallel run where divisions will be sending data to CANSIM I, which will be copied to CANSIM II, and new data series will only be available on CANSIM II.

Secondary distributors are a major revenue source for CANSIM. The database will need to provide these distributors with timely updates to their base. The pricing structure and revenue-sharing arrangements have yet to be discussed and decided upon.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

METHODOLOGICAL DOCUMENTATION OF CASIC SURVEYS ON THE WORLD WIDE WEB

M. Bulmer and R. Thomas[1]

ABSTRACT

This paper describes the development of an electronic Social Survey Question Bank. Its aim is to equip users interested in the design and interpretation of surveys with a critical understanding of survey measurement issues and with information of use in improving and standardizing social survey questionnaires. The conceptual structure of the Question Bank is in part hierarchical, with questions and question sequences organized by social science concepts. Other organizing principles and network structures will be recognized and implemented, such as organization by source survey and year of data collection. Issues of user interface and dissemination via the WWW are discussed.

KEY WORDS: Survey questions; Questionnaire design; World Wide Web; Electronic dissemination; Survey documentation; Database design.

1. INTRODUCTION

In his opening address to the conference, Gordon Brackstone issued a challenge. Technology in relation to social research is both an opportunity and a threat, he said, and asked where we are going. This paper is a case study which examines the impact of technology in one particular part of the survey process, that of data collection, and how to convey information to those who are not technical specialists.

The Centre for Applied Social Surveys (CASS) is a Resource Centre of the UK Economic and Social Research Council, run jointly by Social and Community Planning Research and the University of Southampton, with the University of Surrey. It was established in November 1995 for five years in the first instance. It provides short courses in survey methods and is developing an online, electronic, social survey Question Bank for use by social scientists and social researchers in the academic world, government, market research and the independent and voluntary sectors. The latter is the subject of this paper. The Question Bank is a purely electronic resource, and requires access to the World Wide Web via an Internet connection, with suitable browser software, to access this resource. It can be reached at our World Wide Web (WWW) site, whose Universal Resource Locator (URL) is: http://www.scpr.ac.uk/cass/. To reach the Question Bank, go to the above URL and follow the link to the Question Bank.

In this talk we focus on issues raised by the conceptual design and technical implementation of the Question Bank. Our backgrounds are in social science and survey measurement and we have no claim to particular expertise in computing and computer communications. Our aim is not to push forward the frontiers of technology, but rather to find the best way of exploiting existing technology to serve a particular, fairly demanding, purpose. The Question Bank is still very much at a developmental stage, even though obviously some key strategic decisions have been made. This is a discussion of work in progress.

2. THE AIMS OF THE QUESTION BANK

These include the creation of a reference source of information about questionnaires as measuring instruments for social surveys, the encouragement of cross-survey standardization of the questionnaire measurement process in all appropriate cases, and the capture, conversion into electronic form and retrieval by Question Bank users of the questionnaires from major UK public policy surveys, including questionnaire structures and formats and question texts. The Question Bank seeks to provide commentary on conceptual background of measures used in these and other surveys, definitions of terms and concepts, instructions to interviewers, field coding procedures and to disseminate this information to users remotely on their own work stations via the Internet and the World Wide Web.

The cause to which the Question Bank is dedicated is standardization of measurement procedures in quantitative social surveys. It also aims to evaluate critically the existing de facto standards and to suggest ways of testing and improving them. At its most basic, the Question Bank is a reference source for question formats and question wordings as used on major social surveys. The surveys are those which, because of their sample size and design, data collection procedures, permanent and updated nature and official status, provide standards for designing other surveys and for assessing their results. Many are sponsored by the UK government, such as the Decennial Census (treated as a

[1] Martin Bulmer, Academic Director of the Question Bank, Centre for Applied Social Surveys, Social and Community Planning Research, 35 Northampton Square, London EC1V 0AX, and Foundation Fund Professor of Sociology, University of Surrey, Guildford, UK, and Roger Thomas, Director of the Centre for Applied Social Surveys, London, and Director, Survey Methods Centre, Social and Community Planning Research, London, UK.

survey for Question Bank purposes), the General Household Survey(GHS), the Labour Force Survey (LFS), the Family Expenditure Survey (FES), the Family Resources Survey (FRS), the Health Surveys for England and Scotland *etc.* Others are academically driven or otherwise independent of government, such as the British Household Panel Survey (BHPS), the British Social Attitudes Survey (BSA) and the British Election Study (BES).

Nearly all are either continuous or repeated and their topic and question content is periodically reviewed, pruned or extended. Therefore, there is an important extra dimension of changes over time, which extends greatly the potential range of different question forms which might be stored and retrieved through the Question Bank. Between them these surveys set the framework for much applied social, economic, demographic, employment and health research in the UK. Their data collection and empirical measurement methodologies provide models for other surveys and their results provide national standard distributions with which the results of more specialized surveys can be compared. Different reference surveys often have different ways of measuring particular concepts, such as (say) "ethnicity", "household income", "social class" and so on. These differences exist partly because of the differing approaches and priorities of survey sponsors and users and partly for more contingent and practical reasons.

Nevertheless, alignment of measurement instruments and procedures with a national standard source nearly always lends much added power to inference based on surveys and adds greatly to the usefulness of the cumulative stock of empirical findings and data in the relevant area. The price which must be paid to achieve this is a willingness on the part of survey designers to accept a degree of standardization of survey instruments and, in particular of question wordings.

3. WHAT THE QUESTION BANK PROVIDES TO USERS

Our electronic "Question Bank" aims to provide a reference source which users can access remotely on the WWW, using suitable network connections and browser software. They need to be able to locate what they want, to view it and to download it satisfactorily. In addition, they may (or may not) wish to have their attention drawn to background and contextual issues and potential pitfalls which are important when it comes to deciding which measures to incorporate in a survey instrument – or when interpreting the results which that instrument yields.

Some of this paper is concerned with the internal structure of the Question Bank, but there are other issues raised about the electronic dissemination of information. We are proposing to use the Web as the means of delivering information to users, and we assume that all those who wish to use the Question Bank will have access to the WWW, either through their own PC, from a dedicated PC elsewhere in the organization, or through a machine available in an academic library. This machine will have a standard Web graphic browser, though the system we develop must be accessible independently of the browser which is being run.

Such Web 'publishing' raises new issues both for those who are creating the body of information made available for consultation, and for the operation of the Web as a major carrier of relatively technical information, analogous in some sense to a book or manual. At this stage we are tackling the issues of data structure, but at the same time we are aware that there are a number of unresolved issues about how best to present information in this format. These include the ways in which large quantities of text and images can be mounted and retrieved via the Web, how they can be indexed and searched, and how the reader navigates through a body of material not linked sequentially by page numbers, but spatially by links. The analogy of moving around within the space represented by a Rubic cube may be appropriate.

The nature of what is stored in the Question Bank is not conceptually simple. Survey questions are semantic tools of a special kind for conveying meaning between survey designer, interviewer and respondent and for implementing standardized measurement. Individual questions relate to survey topics and have to come to terms with the multifarious attitudes and behavior of people acting in the real world, with all its social, practical and (in the case of government surveys particularly) administrative complexities. In addition, survey questions are embedded in a context of theoretical ideas and assumptions. Underlying each question is a conceptual approach implying a definition of what is to be measured, a measurement rationale and a method of operationalising the required concept or concepts.

Thus for example, the apparently straightforward concept "hours worked" is embedded in a general approach to the measurement of employment circumstances and conditions and labour market behavior. To design an adequate survey question many definitional issues have to be resolved, for example: Is paid overtime to be included? How about unpaid overtime? Does this refer to the person's main job, or does it also include hours worked in subsidiary jobs? If "main job", how is that concept defined? Then there are measurement considerations to be addressed: Is the intention to elicit average daily hours worked during an explicit reference period, or is the concept to be measured "normal hours worked"? Methodological studies have shown that many persons tend to give "stock" answers to this type of question, based on their official terms of employment or what they see as their "normal" behavior: are such answers acceptable?

From this and countless other examples it can be seen that question designers – and also users and interpreters of survey results – need to be aware of a structured context of background knowledge and methodological awareness pertaining to questions on particular topics and of particular types. The Question Bank aims to provide such a context. It will take the form of text notes and other material to supplement question texts and layouts, some of it drawn from published (or semi published) sources, such as methodological annexes and appendices to survey reports, and some of it drafted ad hoc by the creators of the Question Bank, drawing on their own knowledge as methodologists, survey designers, survey analysts and experienced observers of the uses and abuses of survey methods. What we are creating, in effect, is an electronic encyclopedia of social survey question design, structured around the exemplars provided by the standard reference surveys.

4. BUILDING UP THE CONTENT OF THE QUESTION BANK

Given that both the reference surveys individually and survey methodological knowledge generally continue to change and develop, the task of building up the content of the Question Bank is literally unending. Even if no account were taken of future developments, there is a very large backlog of retrospective material to be captured for each major survey that has been going for a long run of years, like the GHS (from 1971), the LFS (from 1973), the FES (from 1957) and the Census (good records go back to 1851). Faced with this rather daunting task specification, we have had to devise and adapt a strategy for building up the content of the Question Bank.

The first principle is that, whatever else it contains a Question Bank, must at least contain question forms in a readily retrieval form. The PDF (Portable Document Format) file has been used as the basic technical method for mounting survey questionnaires in the Question Bank viewed via the Acrobat Reader, a free piece of software. A second principle is that while it is being established the focus will be upon a limited number of major reference surveys. There will be many temptations to deviate from this path. Potential users quite frequently tell us that a specialized Question Bank is just what is needed in their particular field of research and asking when they can expect to see full coverage. We are inevitably having to give rather disappointing answers to some of these enquiries. Our resources are limited and we have to identify and pursue as priority those aspects of Question Bank coverage which will be of the greatest use to the greatest number of potential users. The third principle of our strategy relates to the time dimension. We have elected to start building up the content of the Question Bank from 1991 as a baseline reference year. There are two reasons for this choice. The first is that 1991 is the most recent Census year and the second is that we think most users will be most interested in survey datasets which have relatively recently become available to secondary analysts, via The UK Data Archive at Essex or through other routes where the time-lag is usually of the order of eighteen months to two years. In general, our priority will be to keep up with the more recent datasets, but of course there are also strong arguments for selectively extending the coverage of the Question Bank backward in time, as resources and other priorities permit. Our policy will be to strike a balance acceptable to users between expanding the breadth of coverage and developing the full depth of coverage.

5. THE TECHNICAL BASE OF THE QUESTION BANK

This raises the interesting and, for us critical, technical issue of how well adapted the WWW is to holding large bodies of information analogous to an encyclopedia, and how the structuring of such information can be handled, both on the server and to maximize user-friendliness. The majority of Web sites do not at present attempt to disseminate a great depth of information. Even sites with large number of pages tend to be 'long' and 'thin', with a large number of different items of information spread horizontally. To pursue the Rubic cube analogy, if one is going to attempt to present a more considerable body of textual data which spreads in three dimensions (different survey sources, different types of stored information and different time horizons), what technical issues does this raise for the file structure, and what difficulties may there be in creating the electronic equivalent of an encyclopedia?

6. THE USER INTERFACE

6.1 Navigating the Question Bank

We aim to develop a clear and straightforward design for the Question Bank. At all times it will be possible for the user who may be uncertain about which part of the Question Bank they wish to visit next, to retrace their steps both in terms of returning to the 'home page' or returning to the screen they have just left by using the 'Home' button or the 'Back' button respectively. This aims to give users the confidence to experiment with different methods of accessing the contents of the Question Bank in the certain knowledge that they can always return to a more familiar place if they are feeling lost or unsure about where to proceed to next. The main pages carry a bar allowing the user to return to the Home page of one of the other key pages such as the list of topics, list of questionnaires, or the search engine. The user interface aims to be aesthetically pleasing and informative to the user without offering an information overload.

6.2 Searching the Question Bank

It will be possible to locate a particular question or questions, with related material, in the Question Bank in one of three ways.

(a) It will also be possible to search on a particular survey. If the user knows that they are interested in, for example, the British Social Attitudes survey, they can go directly to that survey.

(b) The installation of the MUSCAT search engine enables the user to search on particular words, the search engine then taking the user to the occurrences of these words either in the HTML or PDF documents on the site, via a list of occurrences which the user can consult.

(c) It will also be possible to search via a general index of twenty-one socio-economic topics, which will leader the user both to conceptual discussion and individual survey questions.

7. THE TRANSITION FROM PAPI TO CASIC AND ITS IMPLICATIONS FOR THE QUESTION BANK

One particular set of technical problems springs from the changing technology of the source surveys. Many of these have converted from traditional paper and pencil interview (PAPI) to computer-assisted methods of data collection, often termed CASIC (Computer Assisted Survey Information Collection). The documentation of these two modes involves, in the first case, the scanning of paper documents and subsequent image processing and, in the second case, the development of standards for documenting sources which exist as computer programmes driving computer screens, rather than as printed questionnaires and other paper documents. Figures 1 and 2 provide two contrasting examples of the difference in appearance of single pages between PAPI and CASIC questionnaires. (They do not convey the differences in terms of sequencing of questions.)

Paper and pencil questionnaires for complex surveys have a number of characteristics in common. They are created by professional survey organizations, and they generally exemplify good questionnaire design practice (important for the Question Bank). Such survey use simple and largely standard semantic conventions to communicate with interviewers, and are divided into labeled topic sections. Question and response texts can be read in interview sequence, and the paper questionnaires indicate skip patterns in an easily understood way. They show instructions to interviewers on how to present questions and tell interviewers how to code and record responses. They indicate how case and document identification are managed, indicate data format through the code-book. Within such questionnaires it is easy to page backwards and forwards, insert bookmarks *etc*.

There are however problems in capturing questionnaire information for the social survey Question Bank, Over the Question Bank time-span, most of the surveys mentioned earlier have converted from PAPI to CASIC. The Question Bank needs to document both modes. In the case of PAPI we can capture paper questionnaires by scanning paper documents and displaying page images on the Web. These images contain the key questionnaire information. But in the case of CASIC, the paper questionnaire "disappears". What can we put in its place? This is a problem both for survey organizations and for the Question Bank

8. CAPTURING CASIC "QUESTIONNAIRE" INFORMATION FOR THE QUESTION BANK

In the UK context, the favored program for CASIC surveys is BLAISE, written for complex public sector surveys by the Netherlands Central Statistical Bureau. This is a relatively complex survey instrument administered from a laptop computer, introducing a new dimension of technical complexity into the survey data collection process. Each questionnaire, which appears as a succession of screens on the laptop containing a succession of questions has been 'written' in BLAISE code by a programmer familiar with the BLAISE terminology, The de facto CASIC programming standard for surveys to be captured for the Question Bank is BLAISE. Some current approaches to documentation are based on automated editing and reformatting of BLAISE code. Others devote effort to line-by-line manual editing, reformatting and rephrasing. The results are illustrated in the right hand side of Figures 1 and 2, and both examples in Figure 3.

In the Question Bank there are special problems as a result. We wish to capture the equivalent of all the methodological information that is available from a PAPI questionnaire, but in the UK CASIC is still unfamiliar to many survey sponsors and users. And very few people outside the large survey organizations understand CASIC programming languages. This is particularly so amongst social science academics who do not often do surveys themselves, but are a target group for the Question Bank. The disappearance of paper questionnaires causes severe "withdrawal symptoms" for such users. They need documentation that they can read, not program code. We have therefore begun to consider how to deal with this problem in the Question Bank.

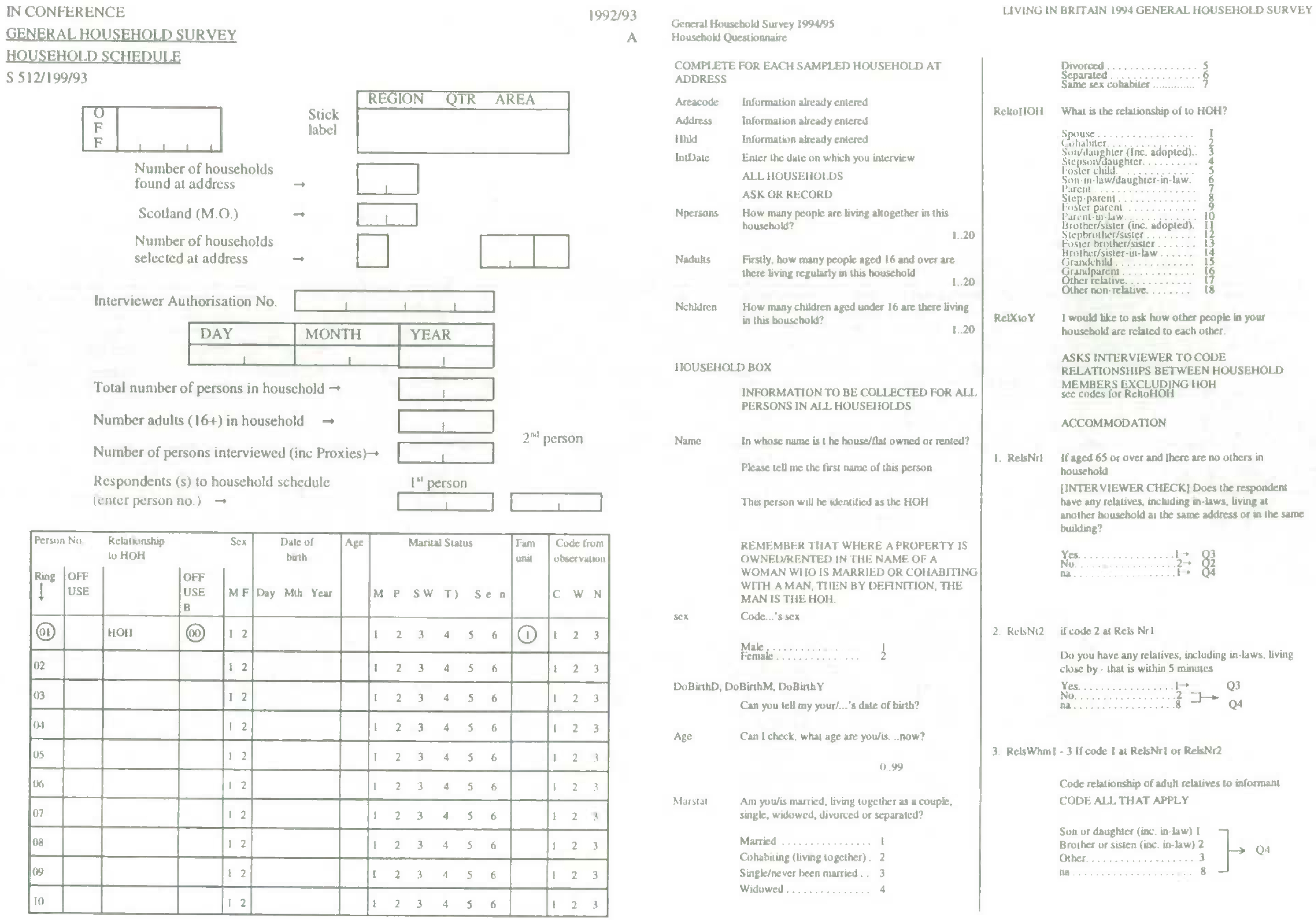
We are reliant upon current approaches to documenting CASIC surveys programmed in BLAISE, some of which are exemplified in Figure 3. All approaches provide question and precoded response texts and definitions of subgroups to which questions are addressed. But there are differing policies on how much concession to make to "plain English" for naive readers, how much to rely on glossaries and preliminary explanation etc to help readers to interpret documentation which is not immediately transparent, and what methodological information to include and what to leave out (*e.g.*, interviewer notes and instructions, online edit checks, *etc*.) No documenters currently provide downloadable program code (at present few users would be able to use this). Some documenters provide illustrative examples of BLAISE screens ("What the interviewer sees").

The Question Bank does not have resources to develop and apply its own CASIC documentation standards. It must rely upon the documentation produced by survey organizations. They often need to communicate with government survey sponsors and users who are unfamiliar with CASIC. Their aims thus overlap with those of the Question Bank, but they are most concerned to convey content and less concerned to convey questionnaire methodology. The introduction of CASIC has created a new situation in relation to the documentation of social surveys the implications of which have not yet been fully assimilated.

9. CONCLUSION

The CASS Question Bank is a new, electronic resource for social scientists interested either in designing questionnaire instruments, or in understanding the detail of the survey measurement process as practiced on a range of major reference surveys. One of its prime aims is to encourage appropriate standardization of survey data collection methods as a means of obtaining more scientific value from the accumulation of survey data. The content of the Question Bank will be built up along three dimensions. The first is the content dimension; the second is the time dimension; and the third is the dimension of conceptual depth and detail.

The Question Bank will stand or fall by its utility to those who access it. The means of access will be via the Internet and the WWW and the processes of capturing the Question Bank material and making it available to users via those media raise many technical problems. In the meantime the development of technology and software in this area is advancing very rapidly, offering potential solutions to problems and also the possibility of offering new Question Bank products and services. The Question Bank user interface needs to include user-friendly aids to navigating the information base and locating question forms and other information relevant to the user's needs.



IN CONFERENCE

GENERAL HOUSEHOLD SURVEY

HOUSEHOLD SCHEDULE

S 512/199/93

1992/93

A

O F F

Stick label

REGION	QTR	AREA

Number of households found at address →

Scotland (M.O.) →

Number of households selected at address →

Interviewer Authorisation No.

DAY	MONTH	YEAR

Total number of persons in household →

Number adults (16+) in household →

Number of persons interviewed (inc Proxies)→

Respondents (s) to household schedule (enter person no.) →

1st person

2nd person

Person No.		Relationship to HOH		Sex	Date of birth	Age	Marital Status	Fam unit	Code from observation
Ring ↓	OFF USE		OFF USE B	M F	Day Mth Year		M P SW T) Sen		C W N
01		HOH	00	1 2			1 2 3 4 5 6	1	1 2 3
02				1 2			1 2 3 4 5 6		1 2 3
03				1 2			1 2 3 4 5 6		1 2 3
04				1 2			1 2 3 4 5 6		1 2 3
05				1 2			1 2 3 4 5 6		1 2 3
06				1 2			1 2 3 4 5 6		1 2 3
07				1 2			1 2 3 4 5 6		1 2 3
08				1 2			1 2 3 4 5 6		1 2 3
09				1 2			1 2 3 4 5 6		1 2 3
10				1 2			1 2 3 4 5 6		1 2 3

General Household Survey 1994/95
Household Questionnaire

LIVING IN BRITAIN 1994 GENERAL HOUSEHOLD SURVEY

COMPLETE FOR EACH SAMPLED HOUSEHOLD AT ADDRESS

Areacode Information already entered

Address Information already entered

Hhld Information already entered

IntDate Enter the date on which you interview

ALL HOUSEHOLDS

ASK OR RECORD

Npersons How many people are living altogether in this household?
1..20

Nadults Firstly, how many people aged 16 and over are there living regularly in this household
1..20

Nchldren How many children aged under 16 are there living in this household?
1..20

HOUSEHOLD BOX

INFORMATION TO BE COLLECTED FOR ALL PERSONS IN ALL HOUSEHOLDS

Name In whose name is t he house/flat owned or rented?

Please tell me the first name of this person

This person will be identified as the HOH

REMEMBER THAT WHERE A PROPERTY IS OWNED/RENTED IN THE NAME OF A WOMAN WHO IS MARRIED OR COHABITING WITH A MAN, THEN BY DEFINITION, THE MAN IS THE HOH.

sex Code...'s sex

Male 1
Female 2

DoBirthD, DoBirthM, DoBirthY

Can you tell my your/...'s date of birth?

Age Can I check, what age are you/is. ..now?

0..99

Marstat Am you/is married, living together as a couple, single, widowed, divorced or separated?

Married 1
Cohabiting (living together) . 2
Single/never been married . . 3
Widowed 4
Divorced 5
Separated 6
Same sex cohabiter 7

ReltoHOH What is the relationship of to HOH?

Spouse 1
Cohabiter. 2
Son/daughter (Inc. adopted).. 3
Stepson/daughter. 4
Foster child. 5
Son-in-law/daughter-in-law. 6
Parent 7
Step-parent 8
Foster parent. 9
Parent-in-law. 10
Brother/sister (inc. adopted). 11
Stepbrother/sister 12
Foster brother/sister 13
Brother/sister-in-law 14
Grandchild 15
Grandparent 16
Other relative. 17
Other non-relative. 18

RelXtoY I would like to ask how other people in your household are related to each other.

ASKS INTERVIEWER TO CODE RELATIONSHIPS BETWEEN HOUSEHOLD MEMBERS EXCLUDING HOH
see codes for ReltoHOH

ACCOMMODATION

1. RelsNr1 If aged 65 or over and there are no others in household

[INTERVIEWER CHECK] Does the respondent have any relatives, including in-laws, living at another household at the same address or in the same building?

Yes. . . . . 1 → Q3
No. . . . . 2 → Q2
na . . . . 1 → Q4

2. RelsNr2 if code 2 at Rels Nr1

Do you have any relatives, including in-laws, living close by - that is within 5 minutes

Yes. . . . . 1 → Q3
No. . . . . 2 → Q4
na 8

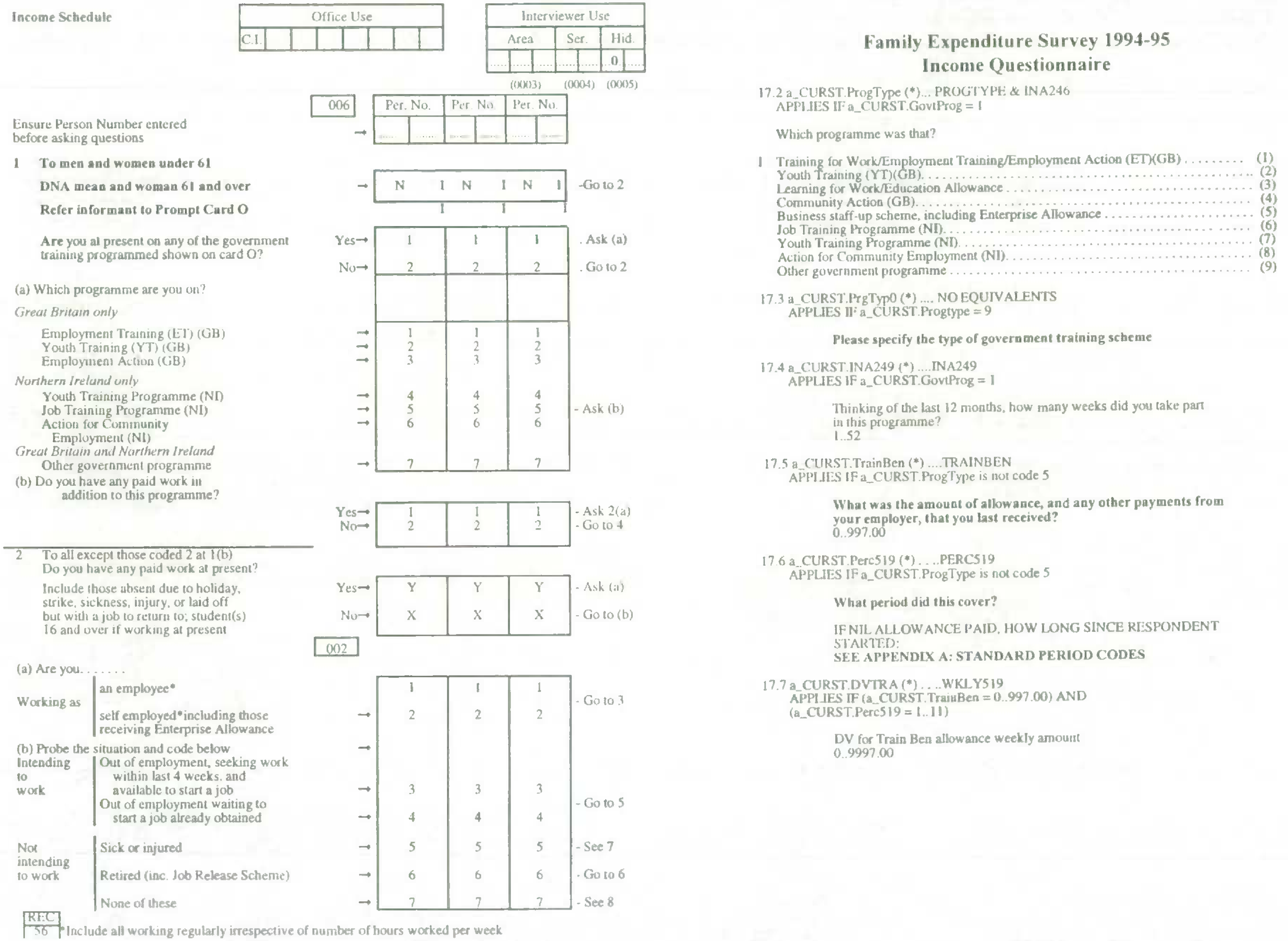
3. RelsWhm1 - 3 If code 1 at RelsNr1 or RelsNr2

Code relationship of adult relatives to informant

CODE ALL THAT APPLY

Son or daughter (inc. in-law) 1
Brother or sisten (inc. in-law) 2
Other. 3
na 8
→ Q4

Figure 1. Comparison of the General Household Survey questionnaire 1992/3 in PAPI and the General Household Survey questionnaire 1994-95 in CASIC (paper equivalent of electronic document)



Income Schedule

Office Use							
C.I.							

Interviewer Use						
Area			Ser.		Hid.	
					0	
	(0003)		(0004)		(0005)	

		006	Per. No.	Per. No.	Per. No.	
Ensure Person Number entered before asking questions	→					

1 **To men and women under 61**					
DNA mean and woman 61 and over	→	N I	N I	N I	-Go to 2
Refer informant to Prompt Card O					
Are you al present on any of the government training programmed shown on card O?	Yes→	1	1	1	. Ask (a)
	No→	2	2	2	. Go to 2
(a) Which programme are you on?					
Great Britain only					
Employment Training (ET) (GB)	→	1	1	1	
Youth Training (YT) (GB)	→	2	2	2	
Employment Action (GB)	→	3	3	3	
Northern Ireland only					
Youth Training Programme (NI)	→	4	4	4	
Job Training Programme (NI)	→	5	5	5	- Ask (b)
Action for Community Employment (NI)	→	6	6	6	
Great Britain and Northern Ireland					
Other government programme	→	7	7	7	
(b) Do you have any paid work in addition to this programme?					
	Yes→	1	1	1	- Ask 2(a)
	No→	2	2	2	- Go to 4

2 To all except those coded 2 at 1(b) Do you have any paid work at present?					
Include those absent due to holiday, strike, sickness, injury, or laid off but with a job to return to; student(s) 16 and over if working at present	Yes→	Y	Y	Y	- Ask (a)
	No→	X	X	X	- Go to (b)
002					
(a) Are you.					
Working as: an employee*		1	1	1	- Go to 3
Working as: self employed*including those receiving Enterprise Allowance	→	2	2	2	
(b) Probe the situation and code below	→				
Intending to work: Out of employment, seeking work within last 4 weeks, and available to start a job	→	3	3	3	- Go to 5
Intending to work: Out of employment waiting to start a job already obtained	→	4	4	4	
Not intending to work: Sick or injured	→	5	5	5	- See 7
Not intending to work: Retired (inc. Job Release Scheme)	→	6	6	6	- Go to 6
Not intending to work: None of these	→	7	7	7	- See 8

REC 56 *Include all working regularly irrespective of number of hours worked per week

Family Expenditure Survey 1994-95
Income Questionnaire

17.2 a_CURST.ProgType (*)... PROGTYPE & INA246
APPLIES IF a_CURST.GovtProg = 1

Which programme was that?

1	Training for Work/Employment Training/Employment Action (ET)(GB)	(1)
	Youth Training (YT)(GB)	(2)
	Learning for Work/Education Allowance	(3)
	Community Action (GB)	(4)
	Business staff-up scheme, including Enterprise Allowance	(5)
	Job Training Programme (NI)	(6)
	Youth Training Programme (NI)	(7)
	Action for Community Employment (NI)	(8)
	Other government programme	(9)

17.3 a_CURST.PrgTyp0 (*) .... NO EQUIVALENTS
APPLIES IF a_CURST.Progtype = 9

Please specify the type of government training scheme

17.4 a_CURST.INA249 (*) INA249
APPLIES IF a_CURST.GovtProg = 1

Thinking of the last 12 months, how many weeks did you take part in this programme?
1..52

17.5 a_CURST.TrainBen (*) ....TRAINBEN
APPLIES IF a_CURST.ProgType is not code 5

What was the amount of allowance, and any other payments from your employer, that you last received?
0..997.00

17.6 a_CURST.Perc519 (*) PERC519
APPLIES IF a_CURST.ProgType is not code 5

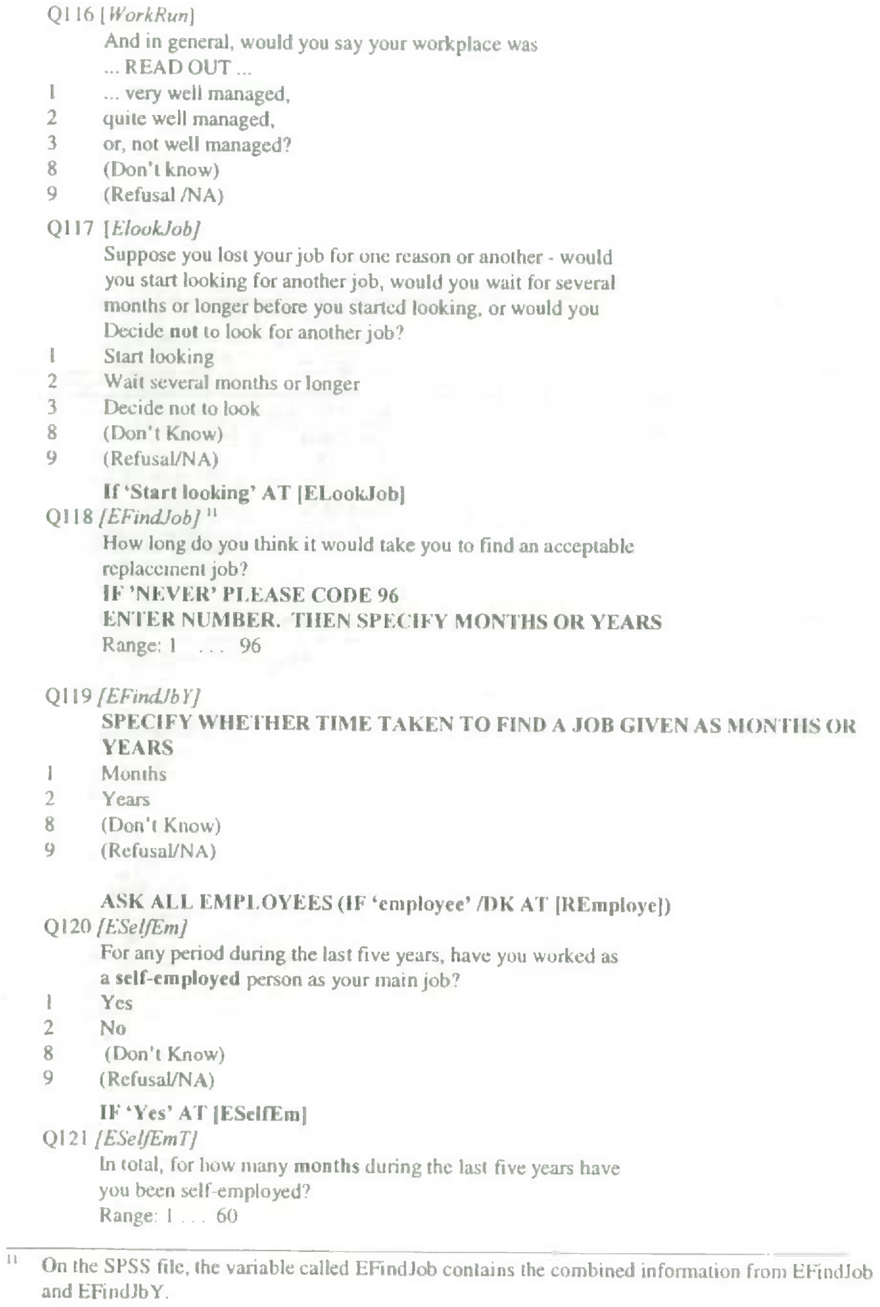
What period did this cover?

IF NIL ALLOWANCE PAID, HOW LONG SINCE RESPONDENT STARTED:
SEE APPENDIX A: STANDARD PERIOD CODES

17.7 a_CURST.DVTRA (*) WKLY519
APPLIES IF (a_CURST.TrainBen = 0..997.00) AND (a_CURST.Perc519 = 1..11)

DV for Train Ben allowance weekly amount
0..9997.00

Figure 2. Comparison of the Family Expenditure Survey questionnaire 1993 in PAPI with the Family Expenditure Survey questionnaire 1994-95 in CASIC (paper equivalent of electronic document).



Q116 *[WorkRun]*

And in general, would you say your workplace was
... READ OUT ...

1 ... very well managed,
2 quite well managed,
3 or, not well managed?
8 (Don't know)
9 (Refusal /NA)

Q117 *[ElookJob]*

Suppose you lost your job for one reason or another - would you start looking for another job, would you wait for several months or longer before you started looking, or would you Decide **not** to look for another job?

1 Start looking
2 Wait several months or longer
3 Decide not to look
8 (Don't Know)
9 (Refusal/NA)

If 'Start looking' AT [ELookJob]

Q118 *[EFindJob]* [11]

How long do you think it would take you to find an acceptable replacement job?

IF 'NEVER' PLEASE CODE 96

ENTER NUMBER. THEN SPECIFY MONTHS OR YEARS

Range: 1 . . . 96

Q119 *[EFindJbY]*

SPECIFY WHETHER TIME TAKEN TO FIND A JOB GIVEN AS MONTHS OR YEARS

1 Months
2 Years
8 (Don't Know)
9 (Refusal/NA)

ASK ALL EMPLOYEES (IF 'employee' /DK AT [REmploye])

Q120 *[ESelfEm]*

For any period during the last five years, have you worked as a **self-employed** person as your main job?

1 Yes
2 No
8 (Don't Know)
9 (Refusal/NA)

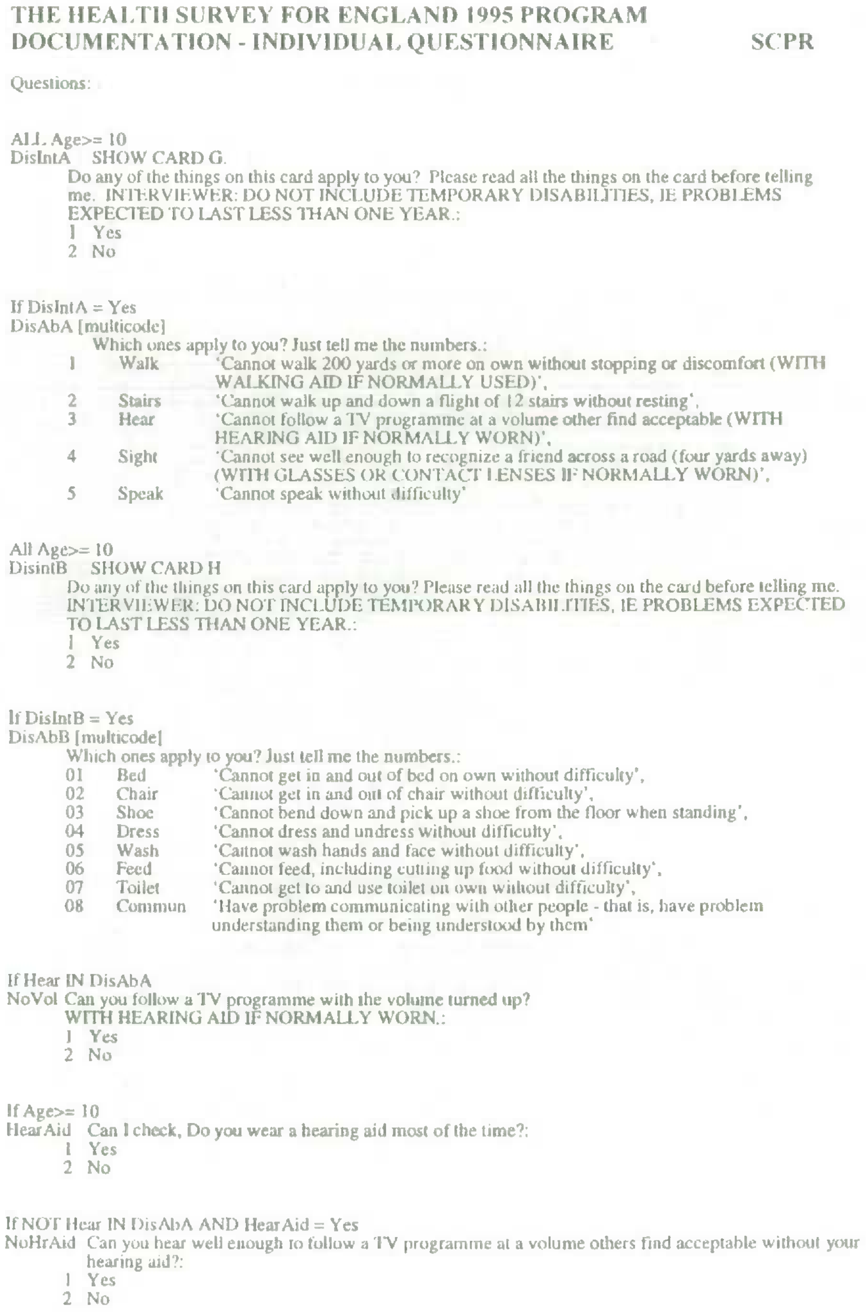
IF 'Yes' AT [ESelfEm]

Q121 *[ESelfEmT]*

In total, for how many **months** during the last five years have you been self-employed?

Range: 1 . . . 60

[11] On the SPSS file, the variable called EFindJob contains the combined information from EFindJob and EFindJbY.



THE HEALTH SURVEY FOR ENGLAND 1995 PROGRAM DOCUMENTATION - INDIVIDUAL QUESTIONNAIRE SCPR

Questions:

ALL Age>= 10
DisIntA SHOW CARD G.

Do any of the things on this card apply to you? Please read all the things on the card before telling me. INTERVIEWER: DO NOT INCLUDE TEMPORARY DISABILITIES, IE PROBLEMS EXPECTED TO LAST LESS THAN ONE YEAR.:

1 Yes
2 No

If DisIntA = Yes
DisAbA [multicode]

Which ones apply to you? Just tell me the numbers.:

1	Walk	'Cannot walk 200 yards or more on own without stopping or discomfort (WITH WALKING AID IF NORMALLY USED)',
2	Stairs	'Cannot walk up and down a flight of 12 stairs without resting',
3	Hear	'Cannot follow a TV programme at a volume other find acceptable (WITH HEARING AID IF NORMALLY WORN)',
4	Sight	'Cannot see well enough to recognize a friend across a road (four yards away) (WITH GLASSES OR CONTACT LENSES IF NORMALLY WORN)',
5	Speak	'Cannot speak without difficulty'

All Age>= 10
DisintB SHOW CARD H

Do any of the things on this card apply to you? Please read all the things on the card before telling me. INTERVIEWER: DO NOT INCLUDE TEMPORARY DISABILITIES, IE PROBLEMS EXPECTED TO LAST LESS THAN ONE YEAR.:

1 Yes
2 No

If DisIntB = Yes
DisAbB [multicode]

Which ones apply to you? Just tell me the numbers.:

01	Bed	'Cannot get in and out of bed on own without difficulty',
02	Chair	'Cannot get in and out of chair without difficulty',
03	Shoe	'Cannot bend down and pick up a shoe from the floor when standing',
04	Dress	'Cannot dress and undress without difficulty',
05	Wash	'Cannot wash hands and face without difficulty',
06	Feed	'Cannot feed, including cutting up food without difficulty',
07	Toilet	'Cannot get to and use toilet on own without difficulty',
08	Commun	'Have problem communicating with other people - that is, have problem understanding them or being understood by them'

If Hear IN DisAbA
NoVol Can you follow a TV programme with the volume turned up?
WITH HEARING AID IF NORMALLY WORN.:

1 Yes
2 No

If Age>= 10
HearAid Can I check, Do you wear a hearing aid most of the time?:

1 Yes
2 No

If NOT Hear IN DisAbA AND HearAid = Yes
NoHrAid Can you hear well enough to follow a TV programme at a volume others find acceptable without your hearing aid?:

1 Yes
2 No

Figure 3. Two examples of CASIC questionnaires (paper equivalent of electronic document), the Labour Force Survey, 1995 and the Health Survey for England, 1995

DEVELOPING AN INTEGRATED HISTORICAL DATA WAREHOUSE FOR EASY ACCESS TO THE REPORTED, SUMMARIZED, AND PUBLISHED DATA OF THE NATIONAL AGRICULTURAL STATISTICS SERVICE

M. Yost and J. Nealon[1]

ABSTRACT

A Historical Data Warehouse is being developed to satisfy the major strategic initiatives in NASS's Strategic Plan. This Data Warehouse will provide integration across surveys and very easy access for all statisticians to multiple years of reporter-level data from our surveys and censuses. This large and specialized data base, which is optimized for rapid ad-hoc query and decision support processing, will have a transformational effect on our survey and estimation procedures. The benefits to our agricultural statistics program that will be realized from our Historical Data Warehouse will be discussed. Our generalized and simple warehouse design schema will be described that will facilitate easy and high speed retrieval of data from different surveys, different time periods, different locations, different sampling units, and other dimensions.

KEY WORDS: Dimensional model; Star schema; Relational database.

1. BACKGROUND

The National Agricultural Statistics Service (NASS) administers the United States Department of Agriculture's program for collecting and publishing timely national and state agricultural statistics. In 1862, as the first Commissioner of the newly formed Department of Agriculture, Isaac Newton established a goal to "collect, arrange, and publish statistical and other useful agricultural information." A year later, in July 1863, the Department's Division of Statistics issued the Nation's first official Crop Production report. The National Agricultural Statistics Service, following the same goal, *collects* data from a population of more than two million farms, and agricultural businesses, *arranges* and summarizes that data into thousands of aggregates and indications, and *publishes* nearly 350 reports a year, covering more than 120 crops and more than 45 livestock items.

For many years, NASS has recognized the critical need for direct access to its historical data for tracking and analysis. Several reports and documents have been published, within NASS, recommending the use of historical data and the need to implement historical databases for analytical purposes. In late 1994, the NASS Strategic Plan formalized a new initiative to develop and implement a Historical Data Base (Data Warehouse) containing census and survey responses, statistical summaries, and official estimates. In doing so, the plan acknowledged the critical role of historical data in improving customer service, reducing respondent burden, expanding analytical capabilities, enhancing sampling and estimation techniques, and improving data quality. What followed was a database designed to track the complete history of data from reporter to official estimates, and to make this rich store of information ***easily*** accessible to ***all*** users within NASS.

2. THE SEEDS OF DATA WAREHOUSING – A BRIEF HISTORY

Like replacing wooden rails with steel because of the Bessemer process, the advent of the direct access storage device (DASD) in the early 1970's marked a turning point in making the data warehouse a possibility. With DASD came new prospects for the storage and access of data, and a new piece of system software called a database management system (DBMS), and a new definition: *database* – a single source of data for all processing. (Building the Data Warehouse, W. H. Inmon, p. 4). These early databases were rectangular and hierarchical and were not very flexible. The data was stored and available, but access was limited to the few who took the time to understand how it was stored and the secrets of retrieval. In the early 1980's, Information Technology (IT) shops began to move from flat file and hierarchical databases to *relational* databases. Cris Date's book on Relational Data Base Management Systems (RDBMS), An Introduction to Database Systems, talks about equal and flexible access to corporate data, and how the RDBMS would answer the problem of turning data into information. There was, interestingly, no discussion on transaction processing or entity/relational modelling. That would come when relational technology would be used to store the transactional and production data of the business. The capture and storage of this data was absolutely necessary and formed the basis of the development of on

[1] Mickey Yost and Jack Nealon, National Agricultural Statistics Service, United States Department of Agriculture, Washington, D.C., U.S.A.

line transaction processing (OLTP). On line analytical processing (OLAP), the knowledge and information counterpart to the OLTP system, would, thus, depend on transactional data stored in data bases optimized for transaction processing. There were, after all, no other sources of data that could be use to understand what was going on in the business.

Just one problem, in order to do transaction processing effectively, data must be modelled in such a way as to permit rapid updating of that data. This translates into a model that is data specific and not at all optimized for storing and querying historical data. Update transactions, under an OLTP model, are made at high rates of speed, but extracts for analysis, under that same model, are very time consuming and slow. The transaction database model uses many join paths against hundreds of tables of similar size and appearance (high normalization) to speed updates. If an end user posed a dimensional query (many joins for much data) against an OLTP data base, (few joins for little data) the OLTP system would drop to its knees. The Data Base Administrator would quickly respond, "you will have to execute that query at night, or on the weekend, or better yet, just let me know what you want, and in a "fortnight" you can have the results!" End users, instead of having direct access to their own data, would have to rely on the IT professional, or highly informed power users, to do ad hoc analysis and reporting from OLTP systems. For the end user, it began to be said, "the price to be paid for the wonderful flexibility of relational databases is that they will always be slow, complex and beyond understanding." This view, fortunately, was wrong.

3. THE STAR JOIN SCHEMA

Enter the star join schema or the dimensional database model. This was what the early RDBMS developers were thinking about, when they described a large central table of numerical facts, surrounded by smaller attendant tables that represented the dimensionality of the facts (a star), see *Attachments 1-3*. The Star Schema represents to the end user a simple and query-centric view of the data by partitioning the data into two groups of data base tables: facts and dimensions. The dimensionality of each fact or data point is completely described by the unique combination of primary keys from each of the attendant dimension tables. In the NASS example, *Attachments 1-3*, the three central tables labelled *Survey Response, Summarized Data, and Official Estimates* each have columns of anonymous index keys, one for each dimension relevant to that table, that relates back to exactly one row in each of the dimension tables. The unique combination of keys in the fact tables is the intersection of each dimensional attribute at that data point. Because the single key in a dimension table is a row, metadata about that key can exist on that row and be available for querying. In fact, the dimensional intersection specified by the unique combination of foreign keys in the fact table defines the data point by presenting the metadata and the data item together. In the NASS application of reporter level atomic data, facts are the numbers recorded in the cells on the data collection instrument, and dimensions are what describe the facts. As can be seen in *Attachment 1*, the NASS dimensions include *TIME, LOCATION, REPORTER, VARIABLE ITEM*, and other dimension tables that serve to partition the facts stored in the *SURVEY RESPONSE TABLE*. The dimensional relationship of attribute information (metadata) to facts is the key reason the star schema is simple and intuitive for the end user. The data are presented in a dimensional, attribute rich format, and organized into the very form needed for analysis, hence the term dimensional analysis. OLTP schemata must be highly normalized into E/R structures for fast transactional updates, and will perform poorly when presented with *ad hoc* queries requiring many rows from many tables. A data warehouse must respond to *ad hoc* queries for information analysis and mining, not transactional updates. The data warehouse built around a star schema clearly has an end user focus, because of the critical need for direct access to key facts using a set of *familiar* and *understandable* dimensions that can be *remembered* over time. The queries are easy to formulate, and ad hoc in nature making analysis easy, which is a key factor in determining whether the warehouse becomes mission critical. The star join schema, therefore, represents the model of choice for on-line integrated data access, from reporter specific data, to official published estimates, organized by dimensions that end users can understand and remember.

It should be noted here that the brief history of database development given above was not the NASS history. NASS did not move its data to RDBMS's in the 1980's, but instead held with mainframe data files that were flat and rectangular. The issues of access and analysis for the end user, however, remained the same: no dimensional access without much programming, and limited direct access to historical data without IT or power user intervention. From now on, for the end user, the strategic initiative of a historical data warehouse, organized around a star schema, would mean a *dimensional data warehouse*, on-line and available to all interested users, and not a collection of flat file data sets stored on magnetic tape and available only by special request.

4. DEVELOPING A STAR MODEL

A group was formed to discuss and specify a dimensional star model. The objective was to ensure the dimensions would be understandable and give end users the opportunity to visualize the data. According to Ralph Kimball, "The ability to visualize something as abstract as a set of data in a concrete and tangible way is the secret to understandability." Obviously a survey time dimension was needed, a reporter dimension for ID's, a location table for counties, districts and states, and a sampling dimension to record stratum descriptions, population counts, and sample sizes. Other dimensions were also proposed, such as a dimension classifying data as keyed, edited, imputed, or revised. Finally, the data column names in the old flat file databases would be re-engineered into a single dimension table. In effect, every column name in the old data structure would become a row in a new dimension table. The old columns were known as varnames. The new dimension table would be named *VARIABLE ITEM* to help the end users realize what was in the table. As the number of rows increased in the *VARIABLE ITEM* dimension (one row for each variable), the number of columns decreased in the fact table, until only a single column of data remained. The *VARIABLE ITEM* dimension table would, therefore, store all the varnames (old column names) and most importantly the metadata describing each of the varnames.

5. METADATA AND THE STAR SCHEMA

The star model is an elegant software solution for organizing and accessing survey metadata. These tables serve the needs of end users by providing, among other things, on-line access to survey and questionnaire specifications, reporter profiling, data classification, and interviewing practices. The metadata is rich and organized visually and in tables that reflect the way the business of the Agency is actually conducted.

Figure 1, is a typical example of a star schema query screen used to return all hog data greater than 0 from *xyz* Farms. With three constraints: ID = "*xyz* Farms", COMMODITY = "Hogs", and DATA > 0, the result set given in Figure 2, was returned.

The columns, YEAR and MONTH from the Time dimension, are in the result set, but were not constrained. With no constraint on the time dimension, the database server returned all months and years *xyz* Farms was in the database. In this example, *xyz* Farms has data from June 1996 and September 1996. Information about the person responding was retrieved from the Respondent dimension as well, without constraint. In June the respondent was the operator, and in September, the hog numbers were estimated because the operator was not available. This leads to a very important result. A query using this schema with as little as two constraints, such as a specific individual and a specific survey, will return an individual's entire questionnaire for that survey, plus the reported data. One implication of this is that extensive previous knowledge about questionnaires, surveys, or specifications is not required to access and understand important information about the Agency and its programs.

6. ISSUES WITH THE STAR SCHEMA MODEL

By converting all of the columns in the old flat file databases into a dimension table, the result was, in effect, a single column fact table with absolutely no sparsity. Every row in the fact table has a meaningful number. Plus, use of the star model is completely generalized. If a new commodity, or a new reporter, or a new stratum definition is needed, all that is required is to add a new row to the appropriate dimension table.

There are problems with this approach, however, because of several issues associated with using a star schema, such as how to handle comparisons among data in the same column. This is not a new problem with the star schema, because of the nature of the tables and their intended use. The dimensions are generally textual and the

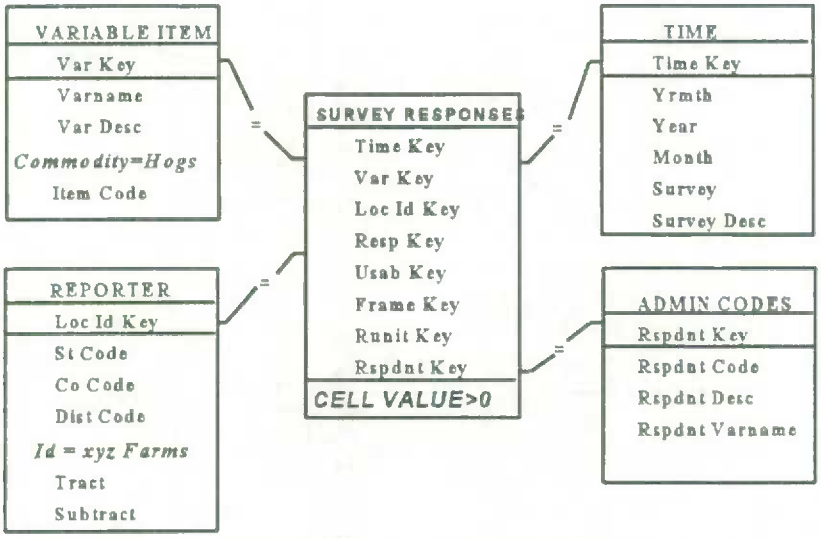


Figure 1. Star Join Schema Used to Access NASS Historical Data With Limits on Commodity, County Code, and All Data

YEAR	MONTH	Rspdnt Desc	Varname	Var Desc	All Data
1996	JUNE	OPERATOR	LHOGBOAR	BOARS & YOUNG MALES FOR BREED	285
1996	JUNE	OPERATOR	LHOGGILT	SOWS, GILTS, & YG GILT FOR BREED	2,715
1996	JUNE	OPERATOR	LHOGTOTL	TOTAL HOGS & PIGS	90,000
1996	SEPTEMBER	ESTIMATED (INACC)	LHOGBOAR	BOARS & YOUNG MALES FOR BREEDING	285
1996	SEPTEMBER	ESTIMATED (INACC)	LHOGGILT	SOWS, GILTS, & YG GILT FOR BREED	2,715
1996	SEPTEMBER	ESTIMATED (INACC)	LHOGTOTL	TOTAL HOGS & PIGS	90,000

Figure 2. Example Result Set Showing Metadata Returned via Cross Reference Through The Fact Table

facts are generally numeric. This tends to create fact tables that are long and narrow as the single column design demonstrates. Not only are there issues with data comparisons, but dimensional ad hoc analysis against a star schema requires multi-table joins between the dimensions and the fact table. These joins cannot be anticipated. Most RDBMS's designed for OLTP will default to a pair-wise join strategy, or the joining of two tables at a time, on all related tables being queried for analysis. Of course, this will probably involve the very large fact table, because it is related to all the other dimensions. This can have a very limiting influence on the analysis if the intermediate result set is too large or must be sampled. There are also issues regarding the referential integrity of the data being loaded into the fact table and the dimension tables. Referential integrity refers to the forced requirement that fact table data must have valid dimension table keys that reference that data back to the dimensions. If data are loaded into a billion row fact table, and there is a missing reference back to the dimension tables that is not enforced at load time, the data item or items loaded into the fact table will be forever lost. There are other important issues in using a star join schema such as indexing strategies, load times, and server segmentation by time periods. Currently these issues are being worked on in the market place.

7. CONCLUSION

Since the first official Crop Production Report, NASS statisticians have grappled with the need to understand the data. There are many influences on the data used to set official estimates for agriculture, and many opportunities for error, both sampling, and non-sampling errors. It is the tracking of these influences and the potential for modelling them against the estimates that gives the data warehouse its true appeal. Every aspect of the business of creating official estimates, from planning, and conducting surveys to statistical methodologies, and data analysis, will be influenced by this new technology. Productive and efficient analysis requires knowledge of the inputs that produce a given output. Data alone does not fulfill this requirement, because it does not carry along the information about the inputs and how they interrelate. This information and knowledge, in the past, has been separated from the data. It may have been available, but only in other disparate data sources, or in manuals and E-mail, or in programs, or in the hip pocket of an analyst. The star join schema represents a relational database model that gathers a great deal of that information and knowledge about the data, stores it, organizes it, and then relates it directly to the factual data being analyzed.

The richness of this information was not available in the transaction models. The emphasis there was on data, not on information. The end user or analyst was dependent on the IT professional or power user to get at the data and report it in such a way that analysis could be performed. If further analysis was required, the process was repeated. The relational star join schema, on the other hand, simplifies the transaction model greatly and is designed for information gathering by the end user. It is an elegant software solution that presents data to the end user in the best possible way to get at the problem of understanding the data.

As information from the data warehouse is used in analysis and decision making, there will be a strong influence on all the processes that create our end product, the official estimates of U.S. agriculture. Operational systems that are choked with both operational and historical data will be freed up to operate more efficiently. As these systems are freed of excess data, re-engineering for efficiencies and quality will be less of a challenge. This re-engineering of tasks and procedures will occur, not because the warehouse needs it that way, but because the warehouse will help uncover data errors and inconsistencies resulting from those tasks and procedures. Perhaps, and most importantly, the information in the warehouse will be used strategically to help carry out the long range goals of the Agency, which is the real reason the data warehouse is a key element in the NASS Strategic Plan.

REFERENCES

Devlin, B. (1997). *Data Warehouse from Architecture to Implementation*, Reading: Addison Wesley Longman, Inc.

Inmon, W.H. (1993). *Building the Data Warehouse*, New York: John Wiley & Sons, Inc.

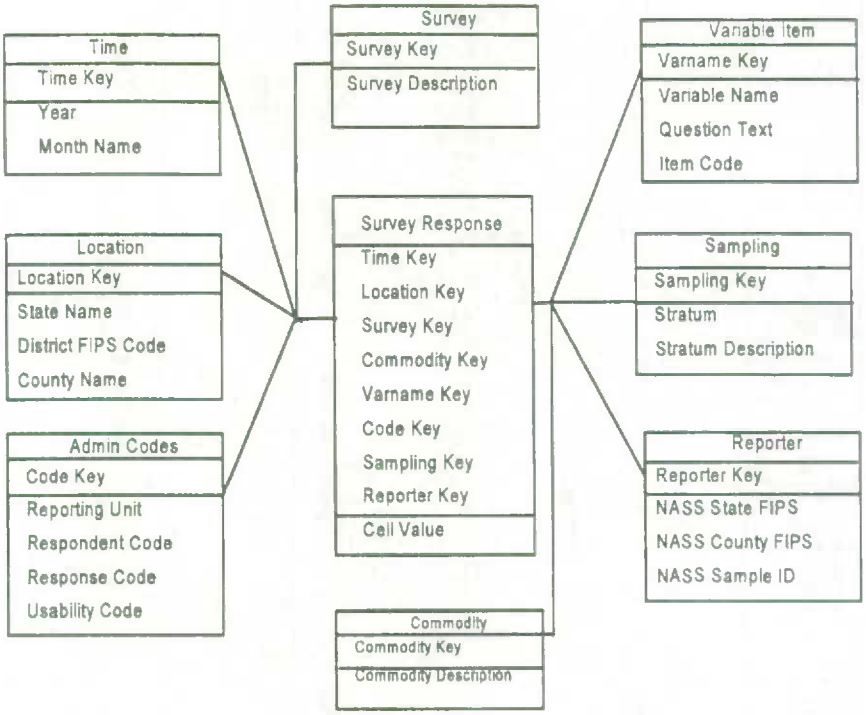
Inmon, W.H. (1997). *Managing the Data Warehouse*, New York: John Wiley & Sons, Inc.

Kimball, R. (1996). *The Data Warehouse Toolkit*, New York: John Wiley & Sons, Inc.

Poe, V. (1996). *Building the Data Warehouse for Decision Support*, Upper Saddle River: Prentice Hall PTR.

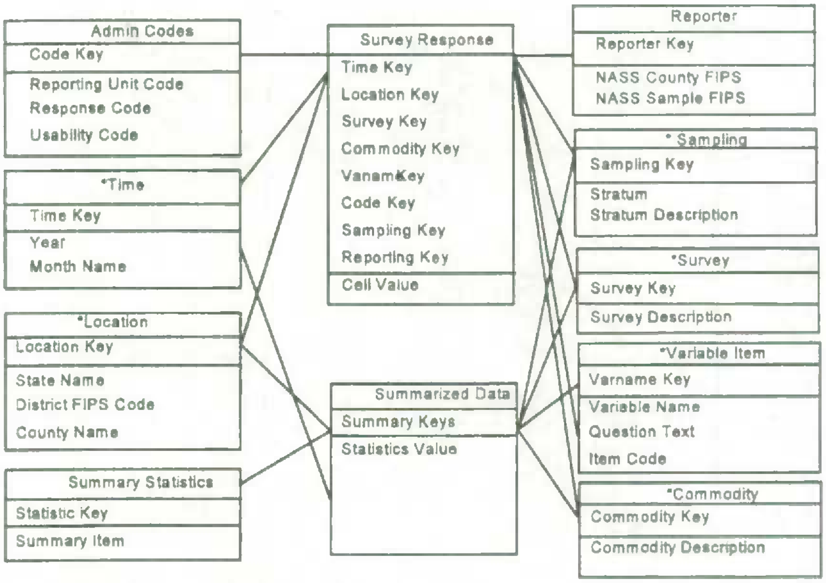
Red Brick Systems, (1996). *Star Schema And STARjoin^{TM} Technology*, Los Gatos: White Paper.

Dimensional Model For Tracking NASS Survey Responses



Attachment 1

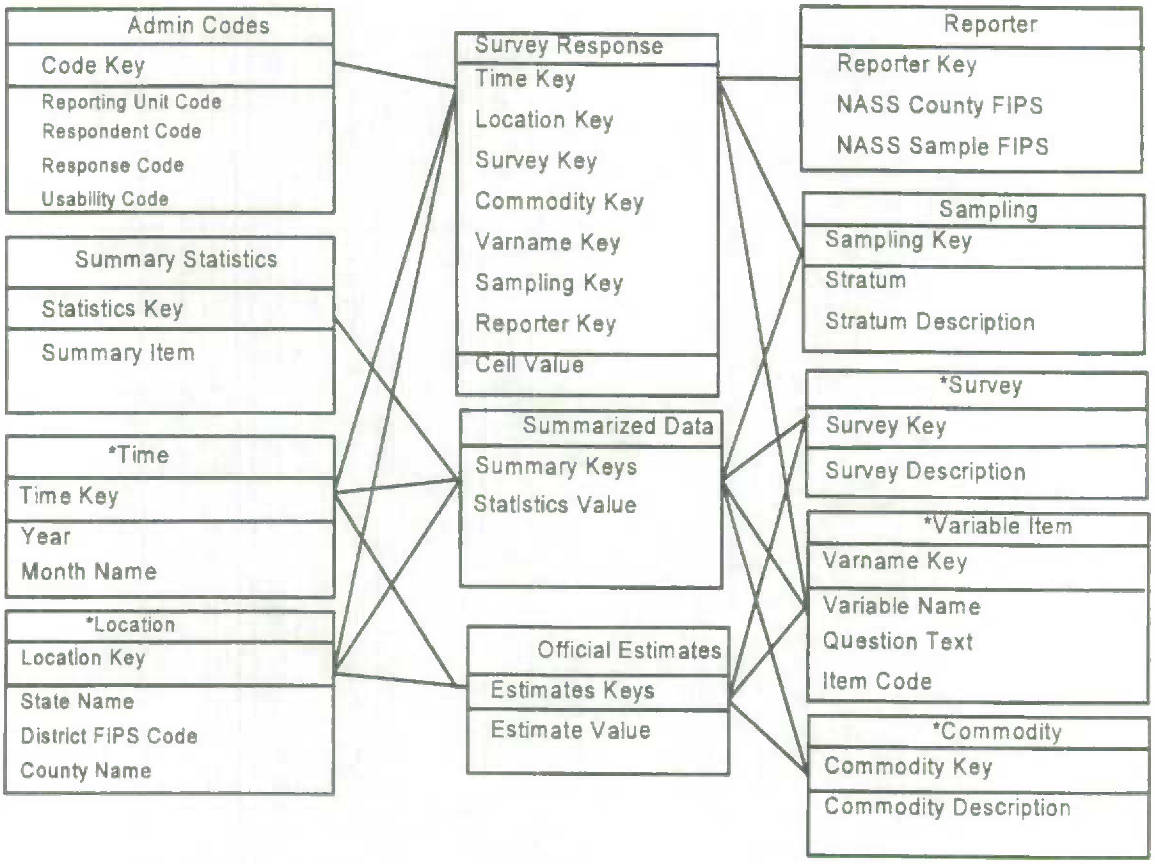
Dimensional Model for Tracking and Linking NASS Survey Responses and Summarized Data



*The Dimension Tables Time, Location, Variable Item, Survey, Sampling , and Commodity are common to both tables: Survey Responses and Summarized Data.

Attachment 2

Dimensional Model for Tracking and Linking NASS Survey Responses Summarized Data, and Official Estimates



*The Dimension Tables Time, Location, Variable Item, Survey, and Commodity are common to all three fact tables: Survey Responses, Summarized Data, and Official Estimates

Attachment 3

SESSION I-7

Geography and Surveys

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

APPLICATIONS OF GLOBAL POSITION SYSTEM (GPS) TECHNOLOGY IN SURVEY TAKING

L. Li and S. Mackie[1]

ABSTRACT

Global Positioning System (GPS) technology enables individuals to capture their location and time, precisely, anywhere on earth. It has potential utility in any survey operation where this information is desired.

This paper provides an overview of how GPS works and its common applications. It also presents results from a feasibility study of how GPS could be used in Canada's 2001 Census to identify the locations of rural dwellings, and describes the use of GPS in a census in Eritrea, Africa.

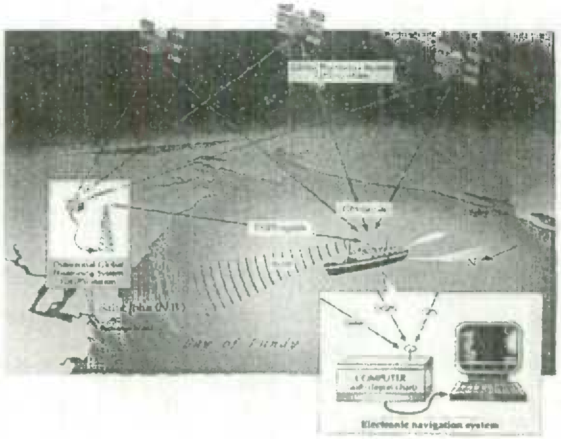
KEY WORDS: Global Positioning System (GPS); Survey taking; Census cartography.

1. INTRODUCTION

Statistics Canada's published reports are the product of numerous operations pertaining to survey design, data collection and processing, and the reporting and publishing of results. In many cases, there is an opportunity to improve the efficiency and quality of these operations through the use of appropriate technology.

This paper introduces the Global Positioning System (GPS) and its role in a variety of survey operations. It is intended to help survey managers understand this rapidly maturing technology and where it may be used to meet survey challenges. Topics covered include:

- how GPS works;
- GPS receivers;
- common uses;
- Statistics Canada's experience; and
- future applications.



2. GPS TECHNOLOGY – HOW IT WORKS

A GPS unit can capture the exact location and time of an event. It receives broadcasts from a constellation of 24 satellites, and triangulates the operator's position from the satellites' signals. Depending on the unit's features, it can also receive and store a variety of accessory information. GPS works under all weather conditions, 24 hours a day, from anywhere with a clear view of the sky.

3. GPS RECEIVERS: TYPES, FEATURES AND COSTS

There are many different models of GPS receivers on the market. They can be sorted into three broad categories: recreation-grade; navigation-grade; and survey-grade.

3.1 Recreation-Grade Receivers

Recreation-grade receivers are pocket-sized, robust, inexpensive ($100 to $500) and easy to operate. Increasingly popular with outdoor enthusiasts, they are often used to guide backcountry travellers, fishermen and hunters. They can identify a location within 5 to 100 metres of its true position.

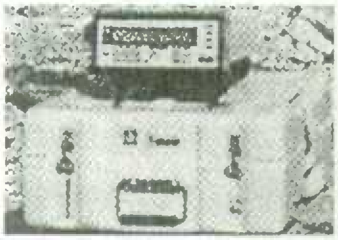
Most recreation-grade receivers can store relatively small amounts of accessory information, and can therefore provide only limited descriptions of each point of location. Similarly, they cannot store enough information about the broadcasting satellites to properly filter signal noise, and therefore cannot consistently be more accurate than 100 meters without additional accessories. However, this may be sufficient for some survey applications such as determining the location of villages, industrial plants, institutional buildings, schools and other large facilities.

[1] Larry Li and Sandra Mackie, Geography Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

3.2 Navigation-Grade Receivers

Navigation-grade receivers are also portable, robust and easy to use. They typically cost between $500 and $5,000. Enlarged data storage capacities enable these receivers to record descriptive information about objects being surveyed and to consistently pinpoint a location within a radius of 3 to 5 metres of its true position. Some of these receivers have been used by Statistics Canada to capture the location of dwellings for census research, and to map footpaths in African villages for census enumeration.

3.3 Survey-Grade Receivers



Survey-grade receivers can store substantial amounts of satellite broadcast data, as well as descriptive data on each feature in the survey. As a result, they are accurate to centimeters of the true position. They have been used by land surveyors to perform legal surveys of properties, by engineers to lay out roadways and sub divisions, and by military personnel to mark the position of mines. Survey-grade receivers cost between $6,000 or more.

Table 1
GPS Receiver Types – Key Features

Receiver Grade	Cost	Accuracy	Features	Uses
Recreation	$100 – $500	5m – 100m	No post-processing capacity	Gross positioning of large facilities, landmarks, backcountry users
Navigation	$500 – $5K	3m – 5m	Data storage for post-processing	Navigation, social surveys, mapping (GIS)
Survey	$6K – $60K	< 1m	Large data storage for post-processing	Legal surveys, precision mapping

4. TRADITIONAL USES OF GPS

Although the use of GPS for social surveys is not widespread, the technology is very mature and has been used extensively in other industries. Parallel situations in statistical survey can be drawn from these established applications, which include:

- facility inventory;
- mapping;
- vehicle navigation;
- marine navigation;
- fleet tracking; and
- land surveying.

4.1 Facility Inventory

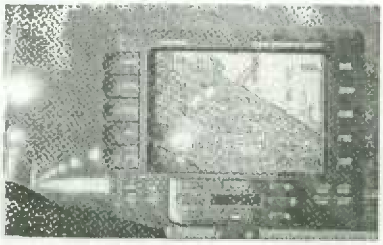
Municipalities and public utilities such as hydro, gas and phone companies use GPS to locate and inventory their assets, such as hydro poles and attached equipment. With GPS, a utility company can easily plan maintenance and determine which parts are required for a specific pole, piece of equipment or area. Such data is also extremely useful for a quick emergency response. From the following photo of a Hydro worker "interviewing" a hydro pole and collecting information about the equipment it supports, the similarities to personal interviewing or computer-assisted personal interviewing are easy to see.

4.2 Mapping

GPS technology can be a very cost-effective tool for mapping. Receivers can capture co-ordinates every second (or every few seconds) and accept descriptive information such as names. This enables an operator to compile detailed data about roads and pathways with little effort, and to map large areas quickly and effectively.

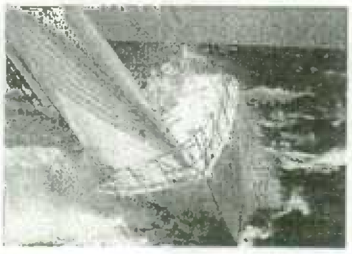
In social surveys, the technology has a useful role within developing nations where it can be difficult and time-consuming to create good base maps to support census and survey operations. Similarly, park managers may find it useful for mapping recreational pathways and planning user surveys.

4.3 Vehicle Navigation



Vehicle navigation is a rapidly growing segment of the GPS market. Today, consumers can buy an electronic road atlas and receiver for approximately $600. The technology runs on a laptop computer. It enables users to pinpoint their location on the street map and obtain suggested routes to their destination. Would similar capabilities be useful for survey interviewers?

4.4 Marine Navigation

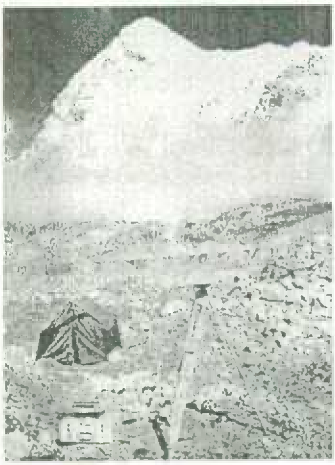


GPS is extremely valuable for navigating in trackless terrain. A large number of sailors and boaters use GPS to monitor their location and plot their routes on traditional navigation charts. In addition, fisheries surveys have employed GPS to link fish catch data in order to improve fish stock management and conser-vation.

4.5 Fleet Tracking

In today's economy, where just-in-time deliveries are required in many manufacturing and transportation industries, companies are using GPS technology to track their vehicles. The systems commonly combine GPS receivers with telecommunications to provide a real-time link between each truck and the dispatch center. This enables the dispatcher to monitor the position and progress of each truck in the fleet, to assign the closest truck for cargo pick-ups, to respond to breakdowns and other unforseen delays, and to communicate anticipated arrival times to the client. Similarly, some police forces are using GPS to monitor the location of police cruisers. It helps ensure that officers closest to a call for assistance are dispatched efficiently, and that any required support units are sent to the correct location. Although real time tracking of interviewers is not often required, similar principals may be applied to tracking the "fleet" of interviewers and the progress of a survey or group of surveys.

4.6 Land Surveying



The extreme accuracy of advanced GPS systems has led to their widespread use by land surveyors seeking to define boundaries for subdivision lots, lay courses for roadways, and locate facilities and engineering works.

The high cost of this technology limits its applicability in social surveys, although it may find a role in agricultural surveys or other small site studies.

5. CURRENT AND POTENTIAL USES AT STATISTICS CANADA

5.1 Feasibility of GPS in the 2001 Census

As discussed, GPS is a fast and precise tool for collecting information on the location of dwellings, business establishments or interviews.

Statistics Canada is evaluating the potential value of GPS technology for the 2001 Census. The tests focus on the feasibility of using GPS to capture the location of dwellings in rural areas. This may improve census operations and allow census data to be tabulated more specifically to meet client needs. It would also facilitate the creation of a Rural Dwelling Register, which would complement the Address Register and provide national coverage to support Statistics Canada's surveys and censuses.

Three tests have already been completed. Together they involved more than forty enumerators and ten thousand dwellings in rural areas outside Ottawa, Toronto and Edmonton. The findings from the most recent test are revealing:

> "The test demonstrated that Census Representatives (CRs) can capture the location of dwellings without seriously impeding or compromising their other duties and responsibilities during the questionnaire drop-off operation. The quality of work performed by the CRs during the test was, in most respects, consistent with the level of work observed in the 1996 Census. Any observable differences in the quality of the work observed on the test can be attributed to factors unique to the test environment or differences in the way the test CRs were trained." (Mark Annett, Regional Census Manager, Prairies & NWT Region)
>
> "It was further concluded that, with some modifications to the GPS receiver, adjustments to the training of Census Representatives on the GPS device and minor changes to the Visitation Record, Census Representatives can capture the location of dwellings, using a GPS receiver, without compromising their other drop-off duties and responsibilities." (Gilles Frechette, Assistant Chief, Survey Operations Division)

5.1.1 Quality of the Data

The quality of data obtained from the tests were very high. Locational co-ordinates were obtained for 92.6% of the 10,244 dwellings in the test. Approximately 95% of the GPS co-ordinates were correctly linked to the household numbers recorded on the EA map.

The position of the dwellings captured by GPS were highly compatible with Statistics Canada's road network, with the dwellings in 38 out of 44 test EAs fitting perfectly. In six EAs, where new roads have been added to the basemap using local maps, the GPS data enabled identification of positional inaccuracies in the roads.

The GPS coordinates captured the occurrence of coverage errors due to misinterpretation of EA boundaries by the enumerators. Instances where enumerators strayed outside their assignment area were evident when the dwelling locations were mapped. This confirms anecdotal evidence from past censuses. With the evidence in hand, field supervisors or survey managers can easily correct the errors to improve the overall quality of the results.

5.1.2 Potential Application

Based on the test results, each of the roughly 4,000 enumerators in rural areas could be given a GPS receiver to use during a census. When dropping off the questionnaires, the enumerator would press a button on the receiver while standing in front of each dwelling to capture its location. Next, its three digit household number would be entered to create a link between the location and the questionnaire/data.

The receivers would be sent to Ottawa, where the data would be processed to create the Rural Dwelling Register. This would include:

- the identification number from the questionnaire for each dwelling (the household number for the Census);
- the co-ordinates for each dwelling;
- the time of visit to each dwelling; and
- the address, or a physical description of the dwelling if an address is not available.

5.1.3 Potential Benefits

Using the information in the Rural Dwelling Register, data could be processed quickly to identify coverage errors. These could be corrected in the field or in the editing phase of data processing. The data would also provide a precise record of enumeration activities, enabling survey managers to better understand enumerator behaviour, time-work and other issues, and to facilitate improvements in operations.

The most significant benefits would be realized in the dissemination of data. Having the exact location of each household would permit census data to be tabulated by any desired geographic area. This would enhance longitudinal analysis by enabling replication of historical dissemination areas. It would also enable fitting the data to school districts, health service areas, sales territories, etc. The housing co-ordinates could also be extremely useful for dispatching ambulances and other emergency services.

5.2 GPS in Census Cartography

Statistics Canada is very active in providing technical assistance to other nations. Currently, Statistics Canada is working with the National Statistics Office in the State of Eritrea (north east Africa) to prepare for their first census. GPS is proving to be a valuable tool for establishing a good base map of the country to support this endeavour.

In developing countries, cartographic preparation often accounts for 60% or more of the overall work involved in conducting a census. In Eritrea this presented a special challenge. Thirty years of war in Eritrea substantially dislocated its people and transportation routes. Therefore, their census could not begin until hundreds of settlements, towns, roads and footpaths were remapped. Roads are sparse and are often little more than paths through villages. Further, most streets and rivers do not have names. Identifying these and other important features depends heavily on recognizing and locating known landmarks such as churches, schools and stores. In light of these obstacles, the process of mapping villages using a compass and manually chaining/drawing the paths would have been extremely time consuming.

5.2.1 Process

To meet this challenge, a team of Geographers from Statistics Canada was asked to conduct a GPS study of the country. This work also involved training Eritreans with the GPS and mapping equipment.

Fourteen survey crews were equipped with GPS receivers to map some seven hundred towns and villages. After just one week of training, these crews of recent high school graduates were able to map the footpaths in the villages by walking along each one with a GPS receiver. The receiver took a reading every five seconds. Lacking house numbers or street names, surveyors stopped in front of key landmarks to record their location and description, which provided the physical features for spatial orientation.

The data collected by the GPS receiver was used to develop the comprehensive base maps of the country required for the Eritrean census. With a data rejection rate less than 3%, the operation was a success.

5.2.2 Results/Evaluation of using GPS in Eritrea

Overall, GPS was a useful survey tool in Eritrea. The GPS receivers were:

- Easy and efficient to use. A surveyor could walk down a village path in minutes with the receiver collecting the data.
- Robust and reliable. Receivers were transported around in the back of a truck through the desert for up to a month at a time. Through the year of operation, no malfunctions were experienced.
- Successful in generating high quality results. GPS has helped the Eritrean statistical office map their country and settlements for their upcoming census, quickly and accurately. This technology has helped generate a flexible information base to support statistical and administrative activities in the country.

Based on these positive results, GPS has strong potential for census mapping in developing countries.

6. CONCLUSION

GPS technology can provide precise times and locations for entities and events in a survey. It is easy to use and cost efficient. The GPS receivers are portable and very robust in the field. This technology could be a useful addition to the spectrum of tools available to any survey manager seeking to acquire precise location or time data.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

THE ADVANTAGES OF GEO-DEMOGRAPHIC STRATIFICATION

M. Zaletel and V. Vehovar[1]

ABSTRACT

The data from different national registers (Central Register of Population, Tax Register, Register of Territorial Units, Register of Housing Units), Census '91 data and some other sources (*e.g.*, Telephone Directory) were merged at the level of enumeration areas for the whole territory of Slovenia. There are about 9,000 enumeration areas with around 65 households each. Census '91 is especially rich source of data: for each house its centroid and altitude are known. Because of very difficult terrain in the country, there is very strong correlation between interviewing costs and altitude of the interviewed household. On top of the above described data, we added some data from our telephone and face-to-face surveys, especially non-response, refusal and non-contacted rates, and the travel expenses of the interviewers. All these data were used to build the model of detailed geo-demographic stratification of the country. The model enables us to conduct more efficient sample designs, to minimize costs of surveys and to reduce non-response rates by adjusting interviewers' training for "more difficult" areas.

KEY WORDS: Response rate; Enumeration areas; Administrative data sources.

1. INTRODUCTION

Data from the national registers basically serve for administrative purposes. Their use for statistical purposes is often not enough explored. In the paper, we show a step which was undertaken at the Statistical Office of the Republic of Slovenia to incorporate data from the register in the efficient way to draw the samples of resident persons and households.

The Central Register of Population itself has already been serving as the primary sampling frame source, yet not all information was effectively used. Further, the data from other registers were also not exploited.

Our first step was to use the data from the Register of Territorial Units. Together with the GIS, the centroids of all sampling units were calculated as well as the average distance from the centre of municipalities. This enables the calculation of the expected travel costs of the interviewer. The average altitude was also attached to every sampling unit. The data from the following sources have also been incorporated: Census '91 data, Central Register of Population (CRP), Database on Employed Persons in Republic of Slovenia (DEP), Telephone Database (TD).

The next step in constructing the database was to attach the non-response data from the official surveys, conducted in the past at the Statistical Office of the Republic of Slovenia. The above constructed information system is basically GIS – System. However, its richness is extremely helpful in constructing the optimal stratification, finding designs with minimal field costs and adopting the frame to non-response problems.

In Section 2 we explain the motivation and the background of the problem. In Section 3 the available data is introduced. In Section 4 we explain the analysis of given data in four steps. Finally the results of analysis and their advantages are shown.

2. MOTIVATION AND BACKGROUND

Almost all samples of official surveys in Slovenia have the same sampling design – they are two-stage stratified samples and primary sampling units are usually enumeration areas. The post-survey adjustment for non-response is also quite similar for most of the surveys: weights are calculated at the level of primary sampling units. If adjustment is done at the level of enumeration areas, perhaps we can also predict the non-response at the level of enumeration areas. Another motivation for this idea are certainly the results from some of the countries (*e.g.*, King 1996) where the division of the country into small areas according to the socio-economic variables was made in advance. Then it was proven that the non-response rates vary across socio-demographic types of areas. We decided to generalise the idea: to build the socio-demographic types of enumeration areas according to the non-response rates achieved in some of the official surveys. This model would enable us to predict the non-response rates for similar surveys in the future.

There are about 14,000 enumeration areas (EA) in Slovenia with 45 households each on average. Unfortunately, some of the EAs are very small or even empty, especially in remote areas. This fact caused a lot of problems in the process of sample designing and selection. In 1996, we merged all small EAs with their larger neighbours. We ended up with 9,872 clusters of

[1] Metka Zaletel, Statistical Office of the Republic of Slovenia, Vozarski pot 12, 1000 Ljubljana; e-mail: metka.zaletel@stat.sigov.mail.si and Vasja Vehovar, Faculty of Social Sciences, University of Ljubljana, Kardeljeva ploscad 5, 1000 Ljubljana; e-mail: vasja.vehovar@uni-lj.si.

enumeration areas (CEA) with an average of 65 households. The problem of small EAs vanished almost completely. Since 1996, the primary sampling units in the majority of official surveys are CEAs.

3. DATA USED TO BUILD GEO-DEMOGRAPHIC STRATIFICATION

Our main sources of data used to build geo-demographic stratification were of course administrative data sources. At the same time we also used survey data to evaluate each step of the analysis of administrative data and to judge the importance of variables involved in the models. In this section we first describe all available administrative data sources, then we describe surveys used for evaluation, and, finally, we introduce variables used in later modelling.

3.1 Administrative Data Sources

All the major administrative data sources available at the Statistical Office of the Republic of Slovenia and some other institutions were used:

- Central Register of Population (CRP),
- Census '91 database,
- Database on Employed Persons in the Republic of Slovenia (DEP),
- Register of Territorial Units (RTU),
- Telephone Database (TD).

At this stage of research, the Taxation Register (kept by the Ministry of Finance) has not been included in the estimations, but when the TR is available, the model will be re-estimated.

Very important point which needs to be stressed here concerns the time distance from the Census '91. All data from the Census are obviously now 6 years old, but we took from the Census mostly data about the dwellings and migrations. The Slovenian population is very stable and only about 2% of population has been moving per year till now. In fact, most of those 2% are migrations within the same towns or villages. The situation about dwellings has not changed much in Slovenia since 1991 since not a lot of new dwellings have been build in-between. We can assume that the Census data are good enough for our purposes.

Surveys

We included the following surveys:

- Labour Force Survey 1994, 1995, 1996 (LFS): this survey was conducted annually in May every year. Sample sizes were approximately 8,000 households per year. The whole field work organisation was very similar from year to year: five follow-ups, advance letters, about 140 free-lance interviewers, face-to-face surveys in PAPI mode. Average length of an interview was 18 minutes. The non-response rates were as follows: 8.9% in 1994, 9.0% in 1995 and 10.1% in 1996.
- Household Budget Survey 1993, 1994, 1995, 1996 (HBS): this survey was also conducted annually in December every year. Sample size in 1993 was 4,500 households, while sample sizes in 1994 – 1996 were about 1,400 households. The field work organisation was similar to that of the LFS, except for the number of interviewers. In 1993 there were 109 free-lance interviewers. In later surveys about 30 interviewers were involved. Average length of interview was about 90 minutes. The non-response rates were as follows: 19.7% in 1993, 17.8% in 1994, 18.0% in 1995 and 34.6% in 1996.
- Household Survey on Energy and Fuel Consumption (HSEFC): the survey was conducted for the first time in Slovenia in May 1997. The sample size was 5,000 households. Sample design for one half of the sample was stratified simple random sampling and for another half was two-stage stratified sampling. Therefore only the results for the second half were used in the analysis presented in this paper. The field work was not organised by the Statistical Office of the Republic of Slovenia as in other surveys, but the organisation of the field work was very similar. The number of interviewers was about 100. Average length of an interview was 23 minutes. The non-response rate was 17,9%.

Variables

First of all we defined five sets of variables, concerning (1) persons, (2) dwellings, (3) households, (4) settlements and (5) clusters of enumeration areas. Then we re-calculated all these variables at the level of clusters of enumeration areas. We started the estimation of the model with the following variables:

Table 1
Sources of Independent Variables

		data source				
Set	variable	CRP	Census	DEP	RTU	TD
1	proportion of children under 15 years	■				
1	proportion of persons over 65 years	■				
1	proportion of employed persons			■		
1	proportion of persons with higher education		■			
2	proportion of privately owned dwellings		■			
2	proportion of dwellings in apartment buildings		■			
2	proportion of weekend or summer houses		■			
3	proportion of farming households		■			
4	proportion of migration for school or work out of the settlement		■			
4	if the settlement is a centre of municipality or not				■	
4	type of settlement				■	
5	density of population				■	
5	air distance from the centre of municipality		■		■	
5	telephone coverage					■

The first idea how to use surveys which were conducted in the past was to take the response rate at the level of enumeration areas. After merging all the data we realised that there are some problems with data from the Household Budget Survey 1993. We were able to define the initial sample size and the responses for each CEA, but that was not the case for the ineligible persons. In every survey we usually experience about 5% of ineligible households because of some differences between dejure and defacto addresses of those persons. After some investigation of the problem we concluded that this problem is equally spread all over the country and that the results for response rate and the completion rate (*i.e.*, the number of responses divided by the number of initial sample) are the same. Therefore we simplified the problem and took the completion rate at the level of CEA.

4. ANALYSIS OF AVAILABLE DATA

As it was explained before, analysis of all available data was performed in a few steps. First, completion rates were computed at the level of CEA for each of the variables from independent sources separately to determine their importance and role in the future designs of surveys. Second, simple linear regression was run to determine the level of importance of each of the variables. Then, correspondence analysis was performed to show the quality of categorisation of available variables. According to the results of all these analyses, all possible types of CEA were the input data for cluster analysis where we clustered several types of CEA to get the final geo-demographic stratification.

4.1 First Step – Computation of Completion Rates

Let us first observe a few figures presenting the dependence of CEA completion rates on selected variables. We calculated general completion rates regardless of the survey.

The general finding of this step was that there existed a dependence of completion rate on most of the analysed variables. There is one very important question appearing: are the presented results survey dependent?

Therefore the completion rates for each of the surveys were computed according to all variables from administrative data sources. Below, some of the results are presented.

We notice that the HSEFC is behaving very differently in comparison with the other two surveys which are very similar. The same picture would be given with other variables which are not shown here. Even before the estimation of the model we can expect that we have to estimate separate models for each of the surveys included. At the same time we can say that the model for the HSEFC will not explain a lot of variability in completion rates. One possible explanation is that the completion rate achieved in the HSEFC was very high. But let us have a look first at the estimation of the models.

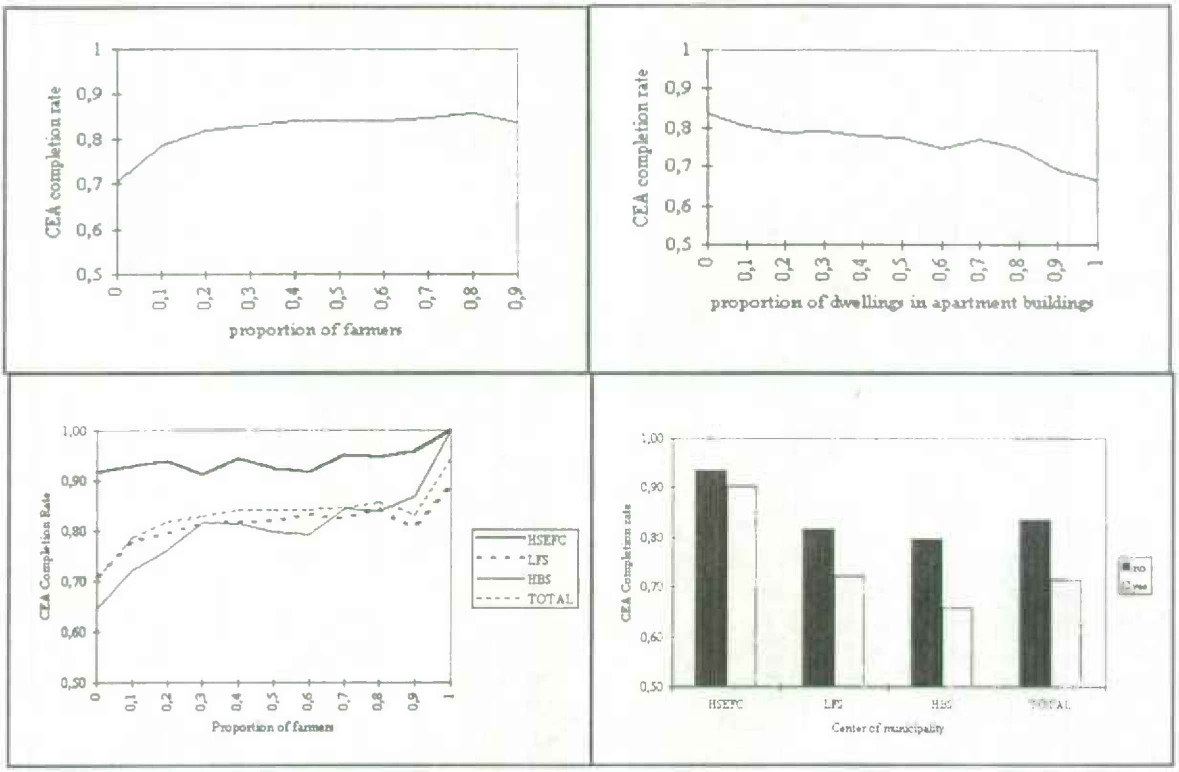


Figure 1. CEA Completion Rate according to some of the independent variables

4.2 Second Step – Linear Regression

The second step was simple linear regression where we wanted to find predictors from available data sources for the completion rate.

The estimation of the regression model has shown what we expected and predicted according to the results of the previous section: the results cannot be generalised independent of the survey topic. Another result seen from the figures above was proved: available variables do not explain the variability in completion rates for the HSEFC at all.

In the table below we labelled only the variables which were significant at least in one of the three regression models. The level of significance is 0.05. Other variables described before are not significant at given level.

Table 2

Linear regression coefficients for three surveys

Set	Variable	HBS	LFS	HSEFC
1	proportion of employed persons	-0.19		
1	proportion of persons with higher education	-0.20	-0.04	
2	proportion of dwellings in apartment buildings	-0.07	-0.11	
3	proportion of farming households		0.02	
4	proportion of migration for school or work out of the settlement	0.03		0.12
4	if the settlement is a centre of municipality or not	0.06	0.03	
4	type of settlement	-0.02	-0.02	
	intercept	1.02	0.86	0.87

We can see that more or less the same variables are significant in the models for the LFS and the HBS. Only one variable is significant for the HSEFC, but even this one does not explain any variability of completion rates. Some of the "demographical" variables are much stronger in the HBS model; on the other hand, "urbanisation" variables are much stronger in the LFS model.

4.3 Third Step – Correspondence Analysis

Almost all used variables were continuous. For simplification and their easier use in the future we decided to define for each of the variables some categories as described below:

Table 3

Break points for the categorization of independent variables

Variable	break point
proportion of children under 15 years	20%
proportion of persons over 65 years	20%
proportion of employed persons	40%
proportion of persons with higher education	40%
proportion of privately owned dwellings	50%
proportion of dwellings in apartment buildings	50%
proportion of farming households	30%
proportion of migration for school or work out of the settlement	30%
density of population	1000 persons/km^2
air distance from the centre of municipality	5 10

Variables not listed in the above table were discrete by their definition or we left them out of the analysis.

The correspondence analysis was computed first of all for three surveys together, then for LFS and HBS together, and finally for HSEFC only. There are two aims of these analyses: first, to evaluate the categorization of variables described before, and second, to compare all three analyses.

Basic results which were produced with correspondence analysis and the comparison of three analyses, are the following: dimension 1 in all three results contains the "urbanisation component" of available variables, dimension 2 in all three results contains more "demographic component". In the case of completion rate in all three results, the first dimension (or the urbanisation dimension) is much stronger than the demographic dimension. Two variables concerning dwellings (proportion of privately owned dwellings and proportion of dwellings in apartment buildings) are in all results approximately the same. In further analysis, we will use only one of them. The effect of demographic variables in the case of the HSEFC is minor in comparison with the LFS and the HBS. Categorisation of variables seems to be reasonable, so further analysis will be run under this assumption.

4.4 Fourth Step – Cluster Analysis

According to the results of the correspondence analysis, we defined all possible types of CEAs. The typology was made according to all available (now discrete) variables. In the first step, we defined them only for "used" clusters. All together, there are all together 382 different types of CEAs; 150 of them occurred only once. This kind of typology is much too detailed for future use, so we decided to analyse all different types with the hierarchical cluster analysis to merge them into reasonable clusters.

5. FINAL GEO-DEMOGRAPHIC STRATIFICATION AND ITS ADVANTAGES

The final geo-demographic stratification was obtained after cluster analysis of different types of clusters of enumeration areas. Forty final types were created on the basis of available variables and the cluster analysis. Types are described with basic properties of CEAs merged together, for example: "families with children, higher education, living in suburban areas around two largest cities".

First of all, final geo-demographic stratification was evaluated according to available data from the surveys. We are interested if completion rates differ across types of CEAs.

We can conclude that this way of stratification is effective also in the prediction of completion rates. Achieved types are very homogenous in distances of CEAs from the centres of municipalities, so the stratification is also effective in the prediction of costs.

6. CONCLUSIONS

In the paper, we show the modelling of detailed geo-demographic stratification in the case of Slovenia. Data from different administrative sources were merged and then analysed according to the results from three official surveys which have been carried out during the last four years at the Statistical Office of the Republic of Slovenia. The aim of building such a detailed stratification was to be able to draw more efficient sample designs, to predict response rates better and to organise field work of future surveys much more efficiently. With correspondence and cluster analysis of all possible different types of clusters of enumeration areas we constructed forty different types of primary sampling units. In the case of completion rates we proved that this kind of stratification is effective.

REFERENCES

Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley.

Groves, R.M., and Couper, M.P. (1993). Correlates of Nonresponse in Personal Visit Survey.

King, J. (1996). The use of geo-demographic coding schemes for understanding household non-response. 7th *International Workshop on Household Surveys Non-response*. Rome, October 1996.

Openshaw, S., Blake, M., and Wymer, C. Using Neurocomputing Methods to Classify Britain's Residental Areas. http://www.geog.leeds.ac.uk /staff/ m.blake /gisruk/gisruk5.html.

Openshaw, S., and Blake, M. Selecting Variables for Small Area Classifications of 1991 UK Census Data. http://www.geog.leeds.ac.uk/staff/m.blake/v-sel/v-sel.html

Vehovar, V. (1993). Field Substitutions in Slovene Public Opinion Survey. *Contributions to Methodology and Statistics*. Ljubljana, FDV, 39-66.

Vehovar, V., and Zaletel, M. (1995). The Matching Project in Slovenia – Who are the Nonrespondents? *Sixth International Workshop on Household Surveys Non-response*. Helsinki, October 1995.

Vehovar, V., and Zaletel, M. (1996). Does the confidentiality concern increase the non-response rate? *7th International Workshop on Household Surveys Non-response*. Rome, October 1996.

SESSION I-8

Advances in Imputation

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

IMPUTING FOR SWISS CHEESE PATTERNS OF MISSING DATA

D.R. Judkins[1]

ABSTRACT

Much of the literature on imputation has focused on univariate imputation problems. However, the most common uses of imputed datasets are multivariate since even the generation of simple cross-tabulations is a form of multivariate analysis. Good simple methods exist for imputing when only univariate analyses will be performed, but methods to support multivariate analyses are necessarily more complex. The only simple multivariate procedure, the common-donor hotdeck is not necessarily any better than a series of univariate imputations. As a result, users have tended to throw out any observations with missing or imputed data when conducting multivariate analyses. A good multivariate imputation scheme should make it possible to both reduce the variance and the bias compared to the method of discarding partially observed records. There has been considerable development in this field using Bayesian approaches to fitting joint models with Monte Carlo Markov Chain algorithms and then using random draws from posterior distributions for the imputations. In this paper, I present a less parametric alternative, called cyclical *n*-partition hotdecks. I also discuss the possibility of cyclic model-based imputations using frequentist models. These methods are contrasted with the Bayesian approach to multivariate imputation in terms of assumptions, suitability for different types of measurements (continuous, categorical, and mixed), ease of implementation, and computing resources. These methods are also contrasted with pure model-based procedures for inference that do not rely on imputation of any sort. Finally, there is some discussion of how to estimate the post-imputation variances with these methods.

KEY WORDS: Hotdeck; Algorithms.

1. INTRODUCTION

We expect some level of item nonresponse in any survey. This complicates the work of all analysts, but particularly the work of secondary analysts who often do not have access to the full set of data about partial respondents nor knowledge of any special circumstances that may have led to the partial response. The data publisher may react to the existence of these difficulties for the secondary analysts in several manners.

The first option is to publish the data as collected. With this option, most secondary analysts will simply drop incomplete cases from their analyses. The set of cases that they drop will depend on the set of variables that they are analyzing. This will cause estimates of the total population with certain characteristics to vary from report to report because different analysts will have chosen to drop different subsets of the sample from their analyses. Another unfortunate consequence of doing nothing is that the estimates will be biased unless nonresponse is unrelated to all the variables being analyzed. This assumption can be made but does not appear plausible in most surveys.

More sophisticated analysts will not be limited to this primitive technique, having the option to conduct maximum likelihood and Bayesian analyses that use models to utilize the partial data. However, these techniques are generally not sophisticated enough to deal simultaneously with the missing data patterns and complex survey designs. As a result, these analyses tend to ignore the clustering and unequal weighting in the design and will generally obtain variance estimates that are too small. New techniques are being developed in this area that can simultaneously account for complex patterns of partial data and for complex sample design, but these techniques are outside the reach of most secondary analysts.

A second option is to only assign weights to cases with no item nonresponse, and to treat the balance of the sample as unit nonrespondents (*i.e.*, either drop them from the published database or give them zero weights). The utility of this approach depends on the percent of cases that have nonresponse to one or more items. If 90 percent of the cases have item nonresponse to a very small number of questions, then this would clearly not be a good approach. Of course, if 75 percent of a questionnaire is blank, then it might be best to define the case as a nonrespondent, but in general it appears desirable to try to retain the partial respondents. In expenditure surveys, in particular, throwing out all partial respondents might result in very small sample sizes.

A third option is to develop a separate set of weights for each anticipated analysis. For example, if it was thought that variables *X*, *Y* and *Z*, were likely to be frequently jointly analyzed (by themselves and in conjunction with other variables), then it might make sense to develop a special set of weights for this set of variables, where only cases that answered all three of the questions were given positive weights and the balance of the sample was treated as

[1] David R. Judkins, Westat Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A.

nonrespondents. Since multiple weights tend, however, to confuse secondary analysts, and since the numbers of unanticipated analyses will be large in a general purpose survey, this approach is probably not the best approach for most surveys.

The fourth option is to impute some or all of the missing data. The attraction of imputation is four-fold. First, imputation makes it possible to have just a single weight for all analyses. Second, it can be made nearly transparent to secondary analysts, making their work very easy, except for the estimation of variance. Third, it preserves more information on the partial respondents than does reweighting. This means that variances are smaller with imputed data than with reweighted data. Fourth, with an imputed dataset all secondary analysts will get the same analyses for margins of tables regardless of whatever variables are used to construct the tables. This property is specifically not shared by the complete-case analysis method since marginal analyses will depend on the set of variables being jointly analyzed.

Although imputation can be done very inexpensively and quickly with simple methods that use a single set of background variables to inform the imputations for all variables with missing values and carry out the imputations independently across variables (such as the single-partition hotdeck), these methods will tend to cause obvious inconsistencies in the data that will lead to the imputed values being ignored by most secondary analysts. This phenomenon is known as attenuation of associations. In its most blatant form, we can imagine finding people with stock dividends but no stocks, post-menopausal women expecting additional children, welfare recipients with substantial real estate holdings and so on. These inconsistencies can happen whenever two variables A and B are associated with each other but either A was not used in deciding how to impute B or vice versa. The secondary analysts of a file imputed through a simple method will generally be wary of attenuation of association and, in order to avoid it, revert to complete-case analyses. Such a user would have been better served by the reweighting solution since the risk of nonresponse biases would have been smaller. Thus, to be useful, imputation should be done with better procedures – procedures that use more of the partial information to maintain observed relationships.

It is fairly easy to design better imputation procedures when the item nonresponse is nested (as was pointed out by Little and David in the early 80s). With nested nonresponse, there is a way of ordering the variables such that if variable $V(n)$ is observed on a particular case, then variables $V(1)$ through $V(n-1)$ are also observed on that case. Nested nonresponse occurs with skip questions and sequential data collection procedures, but it would be surprising to find a large questionnaire to have a strictly nested pattern of nonresponse. Instead, one is likely to find a Swiss-cheese pattern of missing data. Until the advent of inexpensive supercomputing, finding a method to fill in those holes in a manner that did not lead to attenuation of associations was an impossible goal. However, tremendous progress has been made toward meeting this goal over the past decade.

This progress has mostly been due to a group of Bayesian statisticians centered around Don Rubin and Rod Little. In his new book, Schafer (1997) lays out the approach of the Bayesians. Much of the work has been funded by the National Center for Health Statistics as part of their research program for the National Health and Nutrition Examination Survey (NHANES). This work is very good, but it seems to me that there are reasons to try to develop non-Bayesian procedures as well. The primary motivation for other approaches would be to try to reduce computing requirements, to improve the robustness of the methods and to facilitate public exegesis of imputed datasets. We all know that the broad success of surveys in measuring social phenomena is principally due to the work of Morris Hansen and his coworkers at the Census Bureau in the 1940s on design-based inference. This inferential approach is basically nonparametric and frequentist. Confidence intervals are explained as providing information about error levels across a series of hypothetical remeasurements, treating the actual population values as fixed parameters. This approach has led to widespread trust in official statistics. Any approach that relies upon models of superpopulations may face difficulties in becoming widely accepted. Placing prior distributions on the parameters of those superpopulations might further worsen such difficulties.

In this paper, I lay out an approach that is inspired by the work of the Bayesian statisticians but tries to remain more firmly in the nonparametric frequentist tradition. It is not meant as an attack on the Bayesian approach. Rather, the method is suggested as an additional tool that may be employed when there are difficulties obtaining adequate computing resources for Bayesian methods or in gaining trust in the results obtained through the Bayesian methods.

2. METHOD

The method that I am proposing is based upon the hotdeck. There are many variations on the hotdeck. These are discussed in progressive levels of sophistication. With a single-partition hotdeck, a single set of cells is defined that partitions the dataset. Cases with missing values (beggars) are randomly matched with cases with non-missing values (donors) within each cell. The value of the variable for the donor is then copied to the corresponding variable for the beggar. Given a single set of cells (*i.e.*, a single partition), such a hotdeck can be run in less than a week provided that the partition was chosen in advance of examining the data. One problem though with the single-partition hotdeck is that the optimal partition for each variable to be imputed is different. Generally, the partition for a variable is chosen by expert *a priori* judgment about which other variables are most highly predictive of the variables to be imputed. However, the partition can also be chosen through modeling of the variables to be imputed. Modeling produces better results but is more labor intensive.

Thus, an improved approach over the single-partition hotdeck is the n-partition hotdeck, where n is the number of variables to be imputed. Such an approach preserves associations when all the variables that require imputation are conditionally independent of each other given a common core of variables that are completely observed for the entire dataset. In this case, associations between the background variables and the imputed variables will not be attenuated, and there is no association between the imputed variables to attenuate. However, if the variables to be imputed are mutually conditionally dependent given the background variables (as is almost always the case), then an *n*-partition hotdeck will tend to attenuate associations between the imputed variables.

Another option designed to address this problem is the single-partition common-donor hotdeck. With this procedure, a single partition is developed that is a reasonable compromise for all the variables to be imputed. When a donor is matched with a beggar, all the variables to be imputed are copied from the donor to the beggar. This works fine if a reasonable compromise partition can be found and if all the variables to be imputed are block-complete, meaning that they are always reported as a block – either the entire set is reported or the entire set is missing. If the variables are not block-complete, then the two variations are to either copy only those variables that are needed from the donor to the beggar or to overwrite any reported values on the beggar's record with values from the donor. The first approach destroys the ability of the method to preserve associations among the imputed variables while the second approach seems heavy-handed, replacing real data with imaginary data. Furthermore, for a large set of variables to be imputed, the ability to find a reasonable compromise partition for all the variables will not be good. There will generally be more attenuation of association between imputed variables and background variables with a single-partition common-donor hotdeck than with an *n*-partition hotdeck even though the attenuation of association among the imputed variables will not be as bad with a single-partition common-donor hotdeck as with an *n*-partition hotdeck. Neither procedure simultaneously protects associations among imputed variables and associations of imputed variables with background variables. To try to get around this problem, both procedures were applied to different sets of variables on Cycle IV of the National Survey of Family Growth, (Judkins, Mosher and Botman 1991) with reasonable results. Still, there was attrition in some associations with this mixed approach.

Another idea is to impute each variable in multiple runs, imputing a variable on those cases where it is the only missing variable in one run, on those cases where only it and variable *A* are missing in another run, on those cases where only it and variable *B* are missing in another run, on those cases where only it and both variables *A* and *B* are missing in another run, and so on, through all possible patterns of joint missingness. This approach should preserve all types of associations, but is only feasible for small sets of variables because the effort escalates factorially with the number of variables (Ezzati-Rice *et al.* 1993; Winglee, Ryaboy, and Judkins 1993).

It was in trying to resolve this chicken versus egg problem that I noticed the work of the Bayesians. By using Gibbs sampling, they avoided the problem entirely. With that method, it doesn't matter whether one starts with the chicken or the egg, because the initial imputation will be replaced by better imputations in subsequent cycles. It is this cyclic idea that I have borrowed and applied to hotdecks. I like to call the method a cyclic n-partition hotdeck. This method was first described in Judkins, England, and Hubbell (1993) and England *et al.* (1994) in a context with additivity constraints on the set of variables to be simultaneously imputed. With a cyclic method, the variables to be imputed are each imputed a large number of times, using a different partition for each imputation. The basic idea is that if there were only one variable left to be imputed, a partition could be developed for it that would do the best possible job of preserving associations for that variable with all other variables. So, the procedure first imputes something for every variable. It then reexamines each variable sequentially, forming a partition for it based on all the other variables – including the variables with imputed values. As each variable is re-imputed, the best possible value to impute for all the other variables is possibly changed, and so all the other variables are re-imputed. This process continues until some measure of convergence is attained.

This basic idea is simple and appealing, but there are limitations. If the number of variables is large or if any of them are continuous measurements (such as income), then there may be far many more cells than observations. The method breaks down when any cell has a beggar but no donor. For this situation, various *ad hoc* procedures are possible such as reducing the number of variables whose full *n*-way distribution will be protected, categorizing continuous variables when they are used a predictors, and developing procedures for searching for the most similar donor even if it is not in the same cell as the beggar.

3. COMPARISON WITH BAYESIAN METHODS

The Bayesian methods also run into difficulties when there is a large set of variables to be imputed. The Gibbs samplers are generally very slow to converge, particularly when improper or nearly improper priors are used for the variance components. Adding more fixed effects to these models has a strong effect on run times. (The fixed effects in these models are roughly equivalent to the partition used by a hotdeck.) Also, software that I have experimented with such as BUGS from Cambridge University appears to have memory problems if there are more than 20 or so fixed effects.

So the Bayesian methods also require some selectivity in deciding which associations will be protected and which will be neglected. Nonetheless, it does appear that the Bayesian methods can probably protect a larger set of two-way associations than can be done with a cyclic n-partition hotdeck. The hotdeck idea is better suited to the task of protecting higher order interactions among a reduced set of variables. This is easy to see since the hotdeck creates a separate cell for the full Cartesian product of the n background variables. The *ad hoc* procedures for collapsing these cells are generally designed to protect the full set of high order associations of the variable being imputed with the first $n-1$ background variables, totally ignoring the first-order association with the *n*-th background variable. With the Bayesian approach, it is (at least theoretically) possible to sacrifice the higher order associations while continuing to protect all first-order associations.

Another area of major contrast between the Bayesian methods and a cyclic *n*-partition hotdeck concerns continuous predictor variables. The hotdeck may be freely used to impute continuous variables, but there are difficulties in getting it to fully utilize the information in a continuous predictor variable. This is particularly true if there is a strong linear relationship between the variable being imputed (or the logit of a propensity for a target binary variable) and the predictor variable. With the hotdeck, the predictor variable must be categorized and the methods for choosing cutpoints for the categorization are generally *ad hoc*. Of course, if the target variable is only a piecewise linear function of the predictor variable, then the loss from categorization can be quite small.

A third difference between the methods concerns ease of programming. *A* cyclic *n*-partition hotdeck is far easier to program than a Gibbs sampler and thus more certain to succeed within a limited budget and time frame. Programming a Gibbs sampler is fairly easy if all the variables to be imputed are normally distributed, but if there is a mix of continuous, categorical and ordinal variables, the programming is likely to be extremely complex. Furthermore, in order to get the Gibbs sampler to run quickly, there may a need to program it to run on a machine with parallel processing. This can achieve major time reductions since there is considerable work that can be done in parallel for each random effect, but the expertise to design the system is, of course, rare and expensive.

A fourth difference between the methods concerns robustness. Model-based procedures involve assuming a particular shape for the distribution of a variable to be imputed. These shapes are usually simple such as assuming that the mean of the variable is a linear function of other variables and that the errors are identically normally distributed or that the logit propensity of a person to have a particular characteristic is a linear function of other variables and that random effects on the logit scale are normally distributed. Bayesian methods go farther by making this sort of assumption and then also placing a prior distribution on the parameters of the distribution that is assumed to have led to the creation of the realized population. These additional assumptions can be strong or weak, and the failure of the assumptions to hold exactly may or may not be important, but, in general, such assumptions are difficult to justify. Furthermore, when the posteriors are strongly influenced by the priors, the use of these methods exposes the data publisher to questions about motives behind assumptions.

A major advantage of the cyclic *n*-partition hotdeck is that no such assumptions are required. The only assumption that is required for unbiased estimation of first-order statistics such as marginal means is that the mechanism that led to the missing data is ignorable given the partition used for that margin. A similar assumption is also required by the Bayesian methods. The lack of need to assume homoscedastic errors in order to get reasonable imputed values is particularly nice. Consider for example, the problem of imputing total personal financial assets while protecting associations with age and education. The variation in assets is much greater for people in middle age than for young adults. While it is certainly feasible to develop model-based techniques that reflect the variance structure reasonably, more thought and work is required than with a hotdeck.

A particularly interesting violation of the normal error assumption concerns heaping in income distributions. It is well known that respondents tend to round their income off to round figures. This results in heaping in the reported distribution. Model-based imputation will smooth out the heaps since this phenomenon is too complex for the model to capture. A hotdeck approach will replicate the heaped structure in the imputed values. Some might argue that smoothed out values are closer to the truth and thus that we should perhaps go so far as to use the model to replace the reported values. That sort of approach makes me uncomfortable. I think that for government surveys, we should report the data as it was reported to us (with some exceptions for outliers) and that the imputed data should reflect the way the data would have been reported to us had the respondent replied to the question in the same manner as comparable respondents.

Another technique that can be used to make model-based techniques more robust is to add empirical residuals to model-based predictions for missing values, but then the procedure comes to very closely resemble a hotdeck. The only important difference between the two procedures then arises when there assumptions of homoscedasticity fail or when there are very powerful continuous predictors. When the homoscedasticity assumption fails, model-based predictions with empirical residuals can yield values that are outside the allowable range for a variable. Data transformations can reduce this problem, but hotdeck methods always yield plausible values. When there are powerful continuous predictor variables, the hotdeck can lead to slightly greater attenuation of association than a model-based system for variables that are included in the model. This is because, with a hotdeck, the continuous variable must be categorized prior to being used to define the partition.

4. OTHER TECHNIQUES

Model-based frequentist approaches based on maximizing some function resembling a likelihood function are also being developed. The PQL (penalized quasi-likelihood method) popularized by Breslow and Clayton (1993) and implemented in HLM and MLn could be employed in at least three ways. One idea would be to form a separate model for each variable to be imputed for each pattern of missing data on all other variables. With this approach, all observed data would be used to impute the missing values on each variable. However, the variables would still be imputed independently. This would mean that the imputed values across variables might be subject to some attenuation but at least all imputed values would be consistent with all observed data. This idea could be extended in two ways. The first might be to fit a simultaneous model for all variables to be imputed for each separate pattern of missing data. Since this would involve modeling very large covariance matrices, I am somewhat skeptical that this approach could work. The other extension idea would be to use PQL cyclically in the same way that I have suggested using the hotdeck cyclically. This could be done by first using PQL to impute each variable independently. Then use PQL to model the marginal distribution of each variable in turn in terms of all the other variables where the imputed values on the other variables would be treated as if they were observed values.

Comparing such an approach with a Bayesian approach, we might predict that an iterative PQL method is likely to run far faster than a Bayesian approach with the same number of variables. On the other hand, there is also likely to be a negative bias in the estimated components of variance. This means that if the survey is clustered (as most national surveys are), then the imputed values would not show enough variation across clusters. Also, if one were to attempt multiple imputation with such an approach, the variation in the imputed values due to estimating the components of variance would not be reflected, possibly resulting in large underestimates of variance.

Another approach that I have heard of is a variant of the Bayesian method for vectors of categorical variables. The Bayesian approach involves data augmentation and Dirichlet-multinomial priors. Ralph Folsom has an unpublished approach to do something similar with a more frequentist method. His idea is to use a Newton-Raphson type algorithm to solve the likelihood equations where he assumes that the estimators of the regression parameters on the successive logistic scales are jointly normally distributed and then uses sandwich-type variance estimators involving the information matrix to draw whole vectors of imputed values. With this approach, it should be possible to preserve all low-order interactions as in the Bayesian method but without assuming priors. Another advantage of his approach is that it is possible to modify the score equations so that the estimators reflect sampling weights. However, the method does not appear to be extendible to clustered data. He refers to the method as a type of model-based hotdeck.

I mention in passing that Nordbotten's work (Nordbotten 1996) with artificial neural networks seems related, but I haven't been able to study his material closely.

5. VARIANCE ESTIMATION

The design-based variance estimation procedures developed by Rao and other researchers (Rao 1996; Fay 1996a) do not apply to multivariate data with variable missing data patterns. Further improvement of those methods may be possible as Fay (1996b) has predicted, but these methods are not yet available.

Rubin's idea for multiple imputation (Rubin 1996) could probably be used in conjunction with cyclic n-partition hotdecks with reasonably good results provided that certain conditions are met. For multiple imputation to result in consistent variance estimates, it is necessary for the imputation to be proper. The cyclic n-partition hotdeck will almost certainly not be a proper imputation in most applications, but I think that is true for most applications of most imputation methods. There are three major obstacles to achieving proper imputation. The first is that of unplanned analyses. The concept of proper imputation is specific to each analysis. An imputation procedure on a particular dataset may be proper for some analyses and improper for others. Creating a proper imputation for unplanned analyses is probably impossible regardless of the method used. However, if the statistician responsible for the imputation does a good job of anticipating the most important analyses, then the variance estimates obtained through multiple imputation may be pretty good provided that the other conditions are met.

The second obstacle concerns clustering. If clustering introduces a random effect with a nontrivial variance component, then the imputation is only proper when the ratio of the variation in imputed values across clusters to that within clusters is correct. If the hotdeck ignores the clustering, then this will clearly not be the case. With such an approach, there will be a tendency to estimate a negatively biased intraclass correlation. Respecting the clustering, however, means that the partition must have a separate set of cells for each cluster in the sample design. This will generally not leave much room for fixed predictors. Suppose for example, that the statistician is interested in imputing a set of 5 variables, while protecting their associations. This would mean that partition would have to involve the full Cartesian product of 4 of those variables for every cluster in the design. If there are PSUs and second-stage clusters, this will generally not be feasible. Even having a separate partition for every PSU will generally not be possible. This is a point where the Bayesian methods have an advantage since they can borrow strength across the clusters to estimate the fixed effects. It should not be forgotten, however, that fitting large mixed effect generalized linear models with Bayesian procedures is still generally unfeasible. Thus, clustering poses a problem for either approach. Still, multiple imputation may result in acceptable variance estimates provided that the variance components due to clustering are not large – or if they are large, that the partition incorporates the cluster structure.

The third obstacle concerns the size of the donor pool within each cell of the partition. If there is only one possible donor for each beggar, then multiple imputation will dramatically underestimate the variance. In fact, it is likely to do no better than simply treating the imputed values as if they were observed. For the hotdeck imputation to be close to proper, there must be a large pool of donors within each cell. Drawing a bootstrap sample of the potential donors prior to the hotdeck will result in imputations that are closer to proper. Of course, having a large pool of donors is directly at odds with the other obstacles of trying to anticipate unplanned analyses and reflecting the cluster structure in the partition. Solving both of those problems leads the statistician to create a very fine partition with very few potential donors per cell.

Taking these three obstacles into joint consideration, it appears that multiple imputation with the cyclic *n*-partition hotdeck should yield reasonably good estimates when the ratio of sample size to the number of variables is large, when there is good information on likely analyses, and when there is either no clustering or clustering with small components of variance.

6. CONCLUSIONS

Doing a good job of imputing for Swiss-cheese missingness is not an easy task. The cyclic *n*-partition hotdeck is probably the easiest method to implement that has a good chance of keeping attenuation of associations within acceptable levels. Bayesian methods are likely to do better, but may be significantly more difficult to implement well, particularly if the variables have complex nonlinear relationships. Also, the hotdecks are likely to consume far less computer time. Neither method is likely to do a good job for unplanned analyses, neither in the point estimates nor in the variance estimates.

These realizations have led me to consider whether it would not be better to devote our efforts to developing an imputation system that could be easily used by secondary analysts. Such a system could employ either cyclic *n*-partition hotdecks or Bayesian methods depending upon how clever the software designers could be and the capacity of the computers that were assumed to be available to secondary analysts. My idea would be to have a software product where the user would specify the (small) set of variables to be jointly analyzed. The program would then impute the data, possibly incorporating the cluster structure of the design, and possibly performing multiple imputations. The user would then take the custom-imputed dataset into some standard package such as WesVarPC or SUDAAN for analysis.

One objection to such a system might be whether it wouldn't be better to create a software system that directly fits a good model for inference. That seems like an attractive goal, but I think it might be more feasible to achieve a general-purpose imputation package than a general-purpose analysis package for incomplete clustered data. As Don Rubin has pointed out, it requires far less computer time to multiply impute missing values than an entire dataset (replacing observed values with model-based values). It also seems that the analyses would be more robust when only the imputed values rely upon models.

Until such time as a general-purpose imputation package is available, I suspect that the number of survey sponsors who will be willing to pay for careful imputation of Swiss-cheese data will be small. The expenditures for both professional and computer time will tend to be too large to justify when the results only apply to planned analyses. The one exception to this rule may be expenditure surveys. On these, the unit of data collection is frequently a single purchase of some product or service, but the unit of analysis is the person or household that makes the purchases. Since the number of purchases over the course of a year can be large, the percent of sample respondents that remember every detail of every purchase can be quite small. For such a survey, the missing data must be imputed for any analyses of the survey data to be possible at all.

It is thus not surprising that much of the research on imputation has grown out of work on expenditure surveys such as the Medicare Current Beneficiary, the National Medical Expenditure Survey, and the Consumer Expenditure Survey.

On many other surveys such as the Current Population Survey and the Survey of Income and Program Participation, some version of the *n*-partition hotdeck is used although the number of partitions is generally considerably smaller than the number of variables to be imputed. For surveys with this procedure, the secondary analyst must be wary when running crosstabulations or other multivariate procedures. The frequency of rare events such as wealthy people receiving welfare benefits is easily strongly exaggerated with this procedure.

REFERENCES

Breslow, N.E., and Clayton, D.C. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

David, M.A., Little, R.J.A., Samuhel, M., and Triest, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 168-173.

England, A., Hubbell, K., Judkins, D., and Ryaboy, S. (1994). Imputation of medical cost and payment data. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 406-411.

Ezzati-Rice, T.M., Fahimi, M., Judkins, D., and Khare, M. (1993). Serial imputation of NHANES III with mixed regression and hot -deck techniques. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 292-296.

Fay, R.E. (1996a). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fay, R.E. (1996b). Rejoinder. *Journal of the American Statistical Association*, 91, 517-519.

Judkins, D.R., Hubbell, K.A., and England, A.M. (1993). The imputation of compositional data. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 458-462.

Judkins, D., Mosher, W., and Botman, S. (1991). *National Survey of Family Growth: Sample Design, Estimation and Inference*, National Center for Health Statistics, Vital Health Stat. 2(109). (PHS) 91-1386.

Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics*, 12, 385-401.

Rao, J.N.K. (1996). On variance estimation with imputed survey data, *Journal of the American Statistical Association*, 91, 499-506.

Rubin, D.B. (1996). Multiple imputation after 18+ Years. *Journal of the American Statistical Association*, 91, 473-489.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Winglee, M., Ryaboy, S., and Judkins, D. (1993). Imputation for the income and assets module of the medicare current beneficiary survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 463-467.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

FAMILY RESOURCES SURVEY: A PRACTICAL EXAMPLE OF IMPUTATION

J. Semmence[1]

ABSTRACT

The Family Resources Survey (FRS) is the largest survey of household incomes in Great Britain, sampling around 25,000 private households each year. It is sponsored by the Department of Social Security and was launched in October 1992. Participation is voluntary, and the sensitivity and detail of some of the subject matter leads to item non-response, particularly in areas such as income from self-employment and holdings of assets and savings. Following a pilot study at the end of 1993 which looked at a small subset of data from the survey, the decision was taken to develop a full-blown system based on neural network technology. Unfortunately, the FRS neural network system has not lived up to expectations and other methods have had to be tested and employed in its place. The background to the system of imputation which is now being used is covered, before looking at lessons learnt from the FRS experience.

KEY WORDS: Imputation; Neural networks; Household financial survey.

1. INTRODUCTION

The Family Resources Survey (FRS) was launched in October 1992 to meet the information requirements of Department of Social Security (DSS) analysts. The FRS sample aims to cover private households in Great Britain. The target sample size is 25,000 households, interviews being carried out throughout the year. Those interviewed in the survey are asked a wide range of questions about their circumstances. Participation in the FRS is voluntary. Response rates have risen during the life time of the survey from 66 per cent fully cooperating households in 1993-94 to between 69 and 70 per cent for the most recent years.

Fully cooperating households are those where a full interview has been obtained from all adults. However, the definition of a "full" interview incorporates an upper limit of 12 on the number of "don't know" or "refused" answers to key questions (excluding questions on assets). These questions either collect key financial information, such as last pay, or are those which open significant routes. An example of the latter is tenure type where households are routed to a large number of questions on mortgages/rents on the basis of their response.

Overall, the level of missing data is small – less than half a per cent of the more than 10 million set values on the data base are "don't know" or "refused". However, the responses to some questions are much more likely to have missing values than others. Unfortunately, areas of higher non-response tend to be concentrated in variables of particular interest to analysts. Therefore, from a practical stand point, the most important aim of imputation on the FRS has been to maximise the information which is available to users for analysis. Moreover, carrying out imputation centrally simplifies analysis for users and helps ensure consistency of analyses produced using FRS data.

2. THE DECISION TO USE NEURAL NETWORKS

The FRS was drawn to neural networks by the potential benefits on offer:

- It was thought that they might be an opportunity for a quick, automated solution, which would be easy to operate and maintain. Neural networks learn patterns without intervention and although they are heavy on computing resources, they free up analysts' time.
- No information would be wasted: all variables would be used as inputs. The modelling of the complex relationships would make it more likely that the imputed value would be close to the true value and consistent with other values in the case.

In the middle of 1993, a company specialising in neural technology were contracted to undertake a comparative study of neural networks against classical statistical techniques. Two variables were selected for testing. These were weekly premium for structural insurance (continuous) and type of insurance cover, *i.e.*, whether premium paid was for structural insurance, contents insurance or both (categorical). They were chosen by the DSS because they contained data which was considered to be particularly difficult to impute. Using these two variables, the first stage of the pilot compared several different neural network architectures with two alternative imputation methods: replacement by mean/mode and multiple linear regression. These two classical methods were chosen because they were well known, not because they were necessarily thought to be the most effective. The neural network architectures investigated were the Radial Basis Function (RBF); Auto-Associative and Kohonen/Radial Basis Function.

[1] Jo Semmence, Statistician, Department of Social Security, The Adelphi, 1-11 John Adam Street, London, WC2N 6HT, United Kingdom.

The FRS data supplied by the Department consisted of complete records only. First it was examined to identify the different distinct groups present in the data. This process of segmentation is designed to improve the accuracy and distribution of the results. A subset of 1,502 cases for one mortgage, one endowment, no outside contributions was chosen as the data set for imputation within the study. This was then split into training (1,217 records), validation (135 records) and test (150 records) data sets.

During testing, it was found that gains in efficiency could be made if data underwent pre-processing, so that variables had an even spread of information. This was particularly true for the neural network architectures tested, but held for the other methods investigated.

The alternative methods were assessed on the basis of percentage correct values. "Confusion matrices", two way tables comparing imputed and actual results were also examined (for the continuous variable using results formatted into bands), although no formal statistical tests were applied. The results of the study are summarised in Table 1.

Table 1
Percentage of values imputed correctly

Method	Value of weekly insurance premium	Type of structural insurance cover
Replace by mean/ mode	47.3	37.3
Multiple Linear Regression (MLR)/Closest Class mean (CCM)	58.6	52.7
Radial Basis Function	66.7	88.7
Auto-Associative Feed-forward Network	51.3	74.7
Kohonen/RBF Network	84.7	74.7

The conclusion from the first stage of the pilot was that neural networks formed a viable method for the imputation of missing data within the FRS data base, consistently outperforming the other methods investigated. However, during the study several issues had been raised. Although considered difficult, the variables chosen had resulted in relatively high levels of performance from all methods, suggesting data were not as difficult to impute as first thought. DSS therefore requested further tests using additional variables. It was also proposed that work be undertaken to compare different neural network architectures when imputing weakly correlated data.

A second phase to the pilot was therefore instigated, using a similar approach to the first. The same data were used except that sex of the first child in the household (categorical variable) and age first adult left full time education (continuous variable) were merged onto the original set for the weakly correlated data experiments. The variables now considered as difficult were amount of last interest payment on a mortgage and amount of last endowment premium.

For this second phase, tests were repeated on five different splits of the data set into training, validation and test subsets. The results are summarised in Table 2 below and give results averaged over the five tests, as well as the best performance. For weakly correlated categorical data the tests showed that neural approaches still outperformed the statistical methods examined. The final recommendation from the two phases of the pilot study was that a full data pre-processing and imputation system should be developed, based on a Radial Basis Function network.

Table 2
Percentage of values imputed correctly

	MLR/ CCM	Radial Basis Function	Kohonen/RBF Network
Sex			
Average	34.4	41.3	41.9
Best	35.0	43.1	47.1
Age first adult left full time education			
Average	26.0	26.5	
Best	28.0	27.3	
Last interest payment on mortgage			
Average	81.5	82.7	
Best	82.0	86.0	
Last endowment premium			
Average	68.7	70.0	
Best	69.3	74.0	

3. THE FRS NEURAL NETWORK SYSTEM IN PRACTICE

The system was developed to works as follows:

- FRS records are first split into full and partial cases ("full" in this context means without any "don't know" or "refused" responses for any variables in a particular table).
- Full cases are then submitted for training, during which the system first sets up its own validation set and undertakes pre-processing of variables.
- Once training is complete and, where possible, solutions developed, partial cases are submitted for imputation. An option exists to allow replacement by mean or mode where it has not been possible to develop a neural solution. The system provides a solution for any occurrence of a "don't know" or "refused" response in a variable segment it has been told to impute. The output from this process is a series of transaction records which are then applied to the data base.

The system is managed by a series of "control files" for each table of the data base. These are set up in Microsoft Excel and list all variables in the table, followed by the variables to be used for segmentation and then up to 40 additional variables from a maximum of three tables on which the solution is to be based.

Therefore, whether a problem with the technology or the software in which the FRS system has been developed, it does not fulfil two of the promises of neural networks. It does not use all the available information and requires input to set up and decide on which variables to be used (although these do not change substantially from year to year). Nonetheless, even if it were possible to use all variables, it is likely that there would be an unacceptable increase in the time taken for training and any gains in precision would need to be judged against the delays in making data available to users.

4. IMPUTATION OF 1994-95 DATA

The imputation of 1993-94 data was carried out as part of the development of the system. When imputation of 1994-95 data was attempted and output inspected further problems were uncovered:

- At an operational level, the software could not cope with the volume of data from the FRS. Instead of loading data and pressing a button, larger tables had to be split in half and all tables had to be trained and imputed one at a time. This was expected to have an unacceptable effect on the length of imputation, so additional machines were used to allow parallel running (using these additional machines, all stages of imputation were in fact completed in around three to four weeks).
- Only around 20 per cent of all missing values were based on neural solutions and very few were for variables considered particularly difficult to impute. Moreover, around half of these were unusable: either being out of range (the system has no rules to stop this happening) or imputed to the same value (because of the deterministic nature of the networks). The vast majority of results were being set to the mean or mode, since this function had been requested.

That few neural solutions had been found was because it had only been possible to create a small number of networks because of a lack of training data. This resulted from a combination of:

- The complexity of the questionnaire, the routing results in some questions are only asked of a small group of respondents, reducing the availability of information on which to base solutions.
- The requirement that only complete records be submitted for training: the nature of the questionnaire is such that in some tables this is almost impossible to achieve.

All output had to be examined and a decision taken on whether it could be used. All out of range values were automatically deleted. Variables where a large number of cases had been imputed to a small number of values (either mean/mode or neural) were similarly not used. However, time constraints meant that variables with only a small proportion of missing values imputed to a single figure were left. A quick method then had to be found to deal with the remaining missing values. Simple sequential hot-deck algorithms were set up for key variables. Additional methods based on algorithms were also developed.

5. FURTHER TESTING OF NEURAL NETWORKS

It was felt that more testing of neural networks was required, in terms of looking at more variables, more traditional methods and also the effects of imputation on the distributions of variables. To allow comparison with earlier results, the approach was similar to that for the initial pilot. Twenty variables were chosen from the 1993-94 data set, representing categorical and continuous variables which were important in terms of both the extent of missing data and their contribution to key derived variables and analyses. Alternative methods were chosen partly on the basis of what had been used in the past on other surveys. The neural network solutions were created from the DSS system, thereby more closely mimicking what was happening in practice.

In general, the neural network method had higher proportions of correct values for categorical than continuous variables, and results were similar to those achieved previously. However, for continuous variables, results were poorer than for the original test. None of these variables had even half of the imputed values correct to within 10 per cent of the true value. For some variables, such as amount borrowed for a mortgage, only 11 per cent of the imputed data were within a third of the difference of the true value.

Compared with the other methods, neural networks no longer appeared to be the clear front runner, sometimes performing better than traditional methods, other times not. It was also felt that further improvements could be made to the effectiveness of the traditional methods, since only simple solutions (algorithms, number of variables used in regression and hot-decking) had been applied. Nonetheless, from an FRS perspective even these tests were artificial. Like the subsets taken for the original pilot, data were complete for all variables used in the solutions, a situation which is unlikely in practice. For example, despite fairly promising results for the imputation of last pay by neural networks, in practice for 1994-95 no neural solutions were found.

The conclusion from these tests was that no one method stood out at being superior. The most effective method in any particular case depends on the type of variable which is being imputed and the amount and quality of the data being used in the solution.

6. IMPUTATION OF 1995-96 DATA

Armed with the lessons learnt from the additional tests and the experience of 1994-95, a different, more structured, approach was taken for 1995-96 imputation:

- variables with high proportions of missing values and where no neural solutions had been found, or where output had been unusable, were identified and a combination of hot-decking and algorithms used for imputation,
- remaining variables (excluding the table covering Social Security benefits which was imputed as a separate case by case exercise) were then imputed using the neural network system.

The advantage of doing things in this order were that there were more complete data on which to train neural solutions. More variables were also available for segmentation (segmentation variables needing to have no missing data).

Table 3 compares the results between 1994-95 and 1995-96. Hot-decks are now the most frequently used method (60 per cent of all missing values), reflecting the decision to use it to impute variables with the highest proportion of "don't knows" and "refused". Despite more solutions being created, overall, the same proportion of values were imputed using neural networks in 1995-96 as 1994-95 (10 per cent). The contribution of the DSS neural network system to FRS imputation, in terms of the proportion of all values used has decreased from around 40 to 20 per cent of all missing values, although still representing more than 50 per cent of the variables imputed. 417 variables were imputed using the neural network system in 1995-96, 126 neural and 219 mean/mode. 55 variables were hot-decked; 139 imputed using algorithms and 60 were left as missing.

Table 3
1994-95 and 1995-96 imputation compared

	1994-95		1995-96	
Required number of set values in data base	10,409,900		11,279,700	
Missing values prior to imputation	55,300	0.5% of required values	57,100	0.5% of required values
Comparing the Performance of the neural network system				
Neural solutions produced	9,800	20% of missing values	6,000	10% of missing values
Mean/mode solutions produced	34,000	60% of missing values	5,100	10% of missing values
Solutions from RSL system used in final data base	22,200	40% of missing values	10,200	20% of missing values
Comparing methods of imputation				
Neural	6,900	10% of missing values	5,900	10% of missing values
Hot-deck	16,700	30% of missing values	32,100	60% of missing values
Mean/mode solutions	17,000	30% of missing values	4,300	5% of missing values
Algorithms	7,500	15% of missing values	6,600	10% of missing values
Left as missing	7,200	15% of missing values	8,200	15% of missing values

7. CONCLUSION: LESSONS LEARNT

To date, developing an imputation system for the FRS has been a difficult process. We have yet to have a year which runs smoothly: updating our existing system creates new problems and there have always been variables which have to be imputed by less than perfect methods in order to keep to timetable. However, some valuable (and sometimes obvious) lessons have been learnt which may be of relevance to others setting out on a similar task.

Development strategy: As well as testing whether a particular method produces accurate results, it is essential that any piloting should try and mimic the real situation. The FRS pilots took a subset of records for which there were no missing values. Had "real" data been used, that not many neural solutions were developed and that the hardware could not handle the large data set might have been discovered. It might have been possible to find a solution, before the full system was written.

The use of neural networks for imputation: (i) Neural networks need large amounts of data on which to train. As a rule of thumb, the minimum amount of data in a particular segment is $10(M + N)$ where M equals the number of inputs and N the number of outputs. In the FRS case, where networks impute one variable at a time, this amounts to around 500 cases. In a financial survey with specific routing, there are unlikely to be 500 cases for many variables. Overall, for the FRS, results have been disappointing. (ii) They are better suited to the imputation of categorical than continuous variables. (iii) They are not good for imputing variables such as Social Security benefits, for which there are an underlying set of rules which govern amounts.

Imputation: (i) Efficiency of imputation can be improved through pre-processing variables so that there is an even spread of data. (ii) You need to know your data. Imputation for 1993-94 was carried out as part of the development of the system and not by the FRS team. Variables used to segment data did not always make sense, reducing the likelihood of an effective solution being developed. (iii) Users should be consulted. As well as possibly being able to advise on particular regimes, they can also help determine priorities.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

INVESTIGATING NEURAL NETWORKS AS A POSSIBLE MEANS OF IMPUTATION FOR THE 2001 UK CENSUS

M. Cruddas, J. Thomas and R. Chambers[1]

ABSTRACT

In recent years neural networks have been suggested as a possible alternative methodology for imputing missing data. The ONS therefore decided to investigate neural networks as an alternative to the hot deck for imputation in the 2001 Census. Lacking the necessary in-house knowledge of neural networks, Neural Technologies Limited, of Petersfield, Hampshire were chosen, through competitive tender, to work with Census Division of ONS to examine the method. This paper describes that investigation, focusing on the results of an empirical comparison of hot deck and neural network imputation.

KEY WORDS: Hot deck; Imputation; Neural networks.

1. INTRODUCTION

The next UK Decennial Census will be held in 2001. Although every effort will be made at that time to ensure complete data are collected from all census respondents, there will inevitably be a small proportion who do not complete their census forms. These missing data will be imputed at the time of census processing. As part of its development program for the 2001 Census, the UK Census Offices are now evaluating alternative methodologies for this missing data imputation. This paper describes the research undertaken into one such method of imputation, based on the use of neural networks.

The 1991 UK Census enumerated 55.8 million individuals and 22 million households. With 12 person questions and 8 household questions for which imputation was carried out and an average missingness rate of 1 per cent, this represented approximately 8.4 million imputations, and posed significant challenges to software, hardware and the processing timetable at the time.

In the 1991 Census a hot deck process was used to impute missing, invalid or inconsistent observations and it is believed to have worked reasonably well – in the sense of giving sufficiently accurate imputations. However, while it is simple in concept, operationally it is time consuming and resource intensive to develop and program. Furthermore, carrying out modifications to the system at a late stage is often very complicated and slows down the processing of the Census.

In recent years neural networks (NN) have been suggested as an efficient alternative methodology for imputing for missing data (Nordbotten 1996). The ONS therefore decided to investigate NN as an alternative to the hot deck for imputation in the 2001 Census. Lacking the necessary in-house knowledge of NN, Neural Technologies Limited, (NTL), of Petersfield, Hampshire were chosen, through competitive tender, to work with Census Division of ONS to examine the method. This paper describes that investigation, focusing on the results of an empirical comparison of hot deck and NN imputation.

2. WHAT ARE NEURAL NETWORKS?

NN are a class of highly non-parametric regression methods that have proved useful in classification and prediction problems where standard parametric methods are inappropriate (Bishop 1995; Cheng & Titterington 1994). Although originally patterned after the processes assumed to underpin human learning, NN are essentially a class of adaptive statistical methods for computer-based pattern recognition which attempt to emulate (at least in theory) the way humans recognize patterns.

The basic idea behind NN is the creation of a 'fuzzy' mapping from an input data set consisting of a collection of vectors $\{x_1, ..., x_n\}$ to an output data set $\{y_1, ..., y_m\}$, where the input variable x_i represents the information available for classifying the *i*-th item of a 'training set' and the output variable y_i represents the 'true' classification of that item. The mapping is 'fuzzy' because it does not recover the actual value y_i in general. Instead, if we represent the outcome of the NN mapping by the $\hat{y}(x_i)$, then the NN is constructed so that the differences between the $\hat{y}(x_i)$ and the y_i in the training set are "small" according to some appropriately chosen criteria. Once constructed the value $\hat{y}(x_j)$ obtained by applying the NN to a value x_j for which y_j is unknown can then be used to predict this unknown value.

Unlike standard statistical classification algorithms, which typically make assumptions about the conditional distribution of y given x, in order to arrive at an 'optimal' predictor, NN proceed in an essentially nonparametric fashion, building up the classification from the information about the relationship between y and x contained in the training data set. In general, the resulting classifier can be

[1] Marie Cruddas and Jan Thomas, Census Division, Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hampshire, PO15 1RR, UK ; e-mail: marie.cruddas@onsgov.uk. and Ray Chambers, University of Southampton, UK.

represented as a nested sequence of 'coupled' linear and nonlinear transformations of the input data:

$$\hat{y}(x_j) =$$

$$F_{\text{out}}\left(\sum_{k_L=1}^{K_L} u_{k_L} F_{k_L}\left(\cdots \sum_{k_2=1}^{K_2} u_{k_2} F_{k_2}\left(\sum_{k=1}^{K_1} u_{k_1} F_{k_1}\left(\sum_{i=1}^{n} W_{ij} F_{\text{in}}(x_i)\right)\right)\right)\right)$$

That is, an NN consists of an initial transformation of the training data (F_{in}), followed by a set of linear transformations of the resulting values characterized by a set of weights $\{w_{ij}\}$ which depend on the 'new' value x_j, followed by a set of nonlinear transformations (F_{k_1}) corresponding to the first 'hidden layer' of the NN, followed by a second set of linear transformations characterized by weights $\{u_{k_1}\}$ followed by a second nonlinear transformation (F_{k_2}) corresponding to the second hidden layer of the NN and so on. In general there is a sequence of L such hidden layers. At the end, there is a final output transformation F_{out} which gives the NN 'prediction' of y_j.

The number of hidden layers (L), the number and type of transformations involved in the l^{st} hidden layer (K_1), the values of the weights $\{u_{k_1}\}$ used to modify the results of these transformations and the weights $\{w_{ij})$ applied to input data can all (in theory) be determined by a numerical search procedure which attempts to optimize the classifying performance of the NN with respect to the optimality criterion. This process is usually referred to as 'training' the NN. The initial transformation of the input data is usually necessary to ensure that the NN has acceptable operating characteristics (typically referred to as pre-processing), as is the post-processing transformation F_{out} which ensures that the output $\hat{y}(x_j)$ typically is a probability distribution, corresponding to the NN prediction of the conditional density of y_j given x_j and the training data.

The structure of an NN mimics the connectivity of the neuron structure in a human brain, and consequently enables it to identify highly complex nonlinear patterns in the *x-y* relationship. Furthermore, these patterns are identified in a rather automatic way, in the sense that the intrinsic complexity of the NN allows it to represent many more distinct patterns than conventional statistical methods, and consequently provided the training data set is sufficiently rich in such patterns, the numerical optimization methods used to "train" an NN will create "pathways" within the NN for each unique pattern. Finally, because unique *x*-patterns will typically be associated with different values of *y*, these optimization methods allow the identification of those *y* values 'most likely' to arise from a particular *x*-pattern, resulting in the development of a nonparametric estimate of the conditional distribution of *y* given *x*.

Unfortunately, current developments in NN technology do not allow the completely automatic prediction process implied in the preceding paragraphs. A considerable amount of human intervention is still required, in order to specify the types of nonlinear transformations used in pre and post-processing of the training data, and in the specification of the type and number of transformations used in the NN's "hidden layers". However, with the types of hidden layer transformations available at present (*e.g.*, perceptrons, radial basis functions) they are capable of handling quite complex data structures, and so provide a viable alternative to statistical methods for pattern recognition.

3. THE FIRST 'PROOF OF CONCEPT' NEURAL IMPUTATION TRIAL

The objective of the trial, begun in early 1995, was to prove that neural networks could successfully impute missing values in census data. Six variables were chosen for imputation in the trial:

- four "mainstream" variables, **sex** (2 categories), **marital status** (five categories), **number of cars** available to the household (4 categories) and **tenure of accommodation** (eight categories); and,
- two "problem" variables, **age** (single years to 110) and **ethnic group** (nine categories but dominated by one category (White)).

The data supplied to NTL were drawn from five local authority wards (average population of 6,000). The wards covered geographical areas which are known to display different population and housing profiles.

A different neural model was trained for each imputed variable. A model was trained on a particular enumeration district (ED) and an acceptable accuracy level established. If this model was then used for other dissimilar EDs then the level of imputation error increased and the network automatically detected that a new model was needed. The ability of a neural model to impute with accuracy depends on the quality of the data it is trained on: should the demographic nature of the data change significantly, the neural model may become out-dated and will need to be retrained.

The data used in the trial were split into three separate sets:

(a) The '**training set**' - 70% of the data - these were used by the network to generate the model(s) by learning about the relationships between the variables.

(b) The '**test set**' - 20% of the data - these were used to decide when to stop training, thus defining the best generalizing neural model.

(c) The '**validation set**' - 10% of the data - these were the data used to evaluate the performance of the neural models. Effectively, the values of the target variable for 10% of records were blanked out. This was repeated for each variable and then for combinations of variables.

During this first trial the model did not actually provide an imputed value; the results were presented in terms of an **expected probability distribution**. A random number generator would normally then be used to impute a value as a draw from this distribution. As this step was not carried out, evaluation at this stage was concerned with comparing the imputed and actual distributions of the missing data.

The distributions generated by the NN during this trial were judged to be close to those of the missing data, consequently, it was felt that this trial has demonstrated that

NN was a "feasible" approach to imputing missing data in the census (Neural Technologies 1996). A more rigorous follow-up trial was therefore commissioned with NTL to examine the operational and statistical aspects of the method in more detail.

4. THE SECOND NEURAL IMPUTATION 'OPERATIONAL' TRIAL

The aim of the second trial was to examine the feasibility and practicality of using NN in an operational environment and to further examine the statistical properties of the method. Of particular interest was how the process of generating new neural models could be automated in an operational environment. A prototype imputation system was developed by NTL where imputation could be performed on real census data with simulated missing values in order to produce a complete valid dataset as output. This system was then used to "fill up the holes" in a trial data set containing realistic patterns of missingness. This data set was prepared by ONS independently of NTL and provided to them (under strict security provisions), as described below.

The data used to carry out the blind imputation test for the trial was formed by knocking "holes" in an extract of complete data from the post-edit 1991 Census data from two Local Government Districts referred to as County A and County B below. These districts were selected as broadly representative of the UK population.

Table 1
Data used in the second trial

Record type	County A	County B	Total
Households	71,459	69,908	141,367
Private persons	152,878	166,273	319,151

For these complete records, the following two household variables and four person variables were chosen to have 'holes' knocked in them for the NN to then impute:

- Number of rooms	(integer valued)
- Building type	(8 categories)
- Age	(integer valued)
- Marital Status	(5 categories)
- Primary Activity Last Week	(13 categories)
- Country of Birth	(21 categories)

The holes were knocked according to a pre-defined schedule of frequencies which differed according to combinations of variables. The pattern of frequencies was chosen largely according to the patterns observed in the 1991 Census. For example, this resulted in approximately 50% of the data for a variable being deleted in some areas that were considered hard to enumerate in 1991.

Since the value of the variable was not considered when it was deleted, this represented a situation where missingness was at random rather than a possibly more realistic situation where a subset of the population with particular attributes might be more likely to not respond. However this alternative type of missingness is difficult to replicate from complete census data records since, by definition, the value of the items are missing.

5. OPERATIONAL EVALUATION OF THE SECOND TRIAL

The following evaluation is largely based on the report prepared by Dr Jim Austin (University of York) for ONS on operational aspects of the trial NN imputation system supplied by NTL as part of the second trial (Neural Technologies 1997). This system was "built" using the trial data set provided by ONS. It automatically created neural imputation models based upon geographical regions and then carried out multiple field imputation for up to six missing fields.

5.1 Hardware and Software Requirements

The trial NN imputation system was developed in C. It could therefore be implemented on any platform which supports C/C++ code. This includes the ONS Strategic platforms of UNIX and Windows NT. The trial NN system could extract information from the major standard database formats, including Oracle. However, before the data can be shown to the neural model it has to be pre-processed into a form suitable for neural modeling. The choice of database for the 2001 Census is unlikely to create a problem in this regard because data could be extracted into a format suitable for the neural system, as happened in the trial.

5.2 Ability to Handle Census Data Volumes

On the basis of the system developed for the second trial, NTL estimated that the processing power of 74 distributed PCs over two weeks, or 19 distributed PCs over 2 months, would be sufficient to train the networks and impute missing variables in the 2001 Census. These estimates were based on trickle training, *i.e.*, training as data arrives. If the final imputation system trains in batch, NTL estimated that the training would be 3-4 times faster and hence reduce the processing power and time needed. However, by 2001, it is likely that computing performance will have at least doubled, based on recent advances. For scaling up purposes, it was assumed that the 2001 Census will require 76 million person and household records to be processed, and 8 million imputations performed. For these volumes it was estimated that nearly 5 GB of total core data storage space would be required. It was anticipated that this was not likely to be a problem given the current cost and capacity of disc storage.

In comparison, the hot deck system used in the 1991 Census processed data as and when they became available and took five elapsed months to complete the task.

A major problem with NTL's scalability claim was that it was based on a linear relationship between processing power and time taken to process data and the number of records processed. However, the critical aspect of the NN solution is the time taken to train the network. The actual imputation of missing variables takes only about 1% of the training time.

The key factor in determining the processing power needed is the building block used to create the neural imputation models. The NTL estimates above were based on the time taken to build a neural system to impute six variables for two local government districts. The trial data were initially split into small batches, equivalent to wards, and networks were created for each small batch. This initial division was primarily to provide a starting point, so that the model could merge or subdivide these areas if appropriate. If ward is chosen as the imputation "building block" in 2001 and if the wards in the training set are fairly representative, then NTL's estimates probably reflect the processing power required in 2001.

However, neural models are built according to geographic areas and require the setting of threshold parameters. One parameter determines when there is sufficient data in an area to start training, and another determines when the network no longer fits the data and needs to be split into smaller geographic areas. If these parameters are incorrectly set it is possible the system will not model the problem properly, or will require much splitting and retraining, resulting in a general slowing down of the system.

5.3 Impact on Census Processing Flow

The NN imputation system would be a 'black box', standing apart from the normal census processing system, into which all census data would be sent, with complete data being used for training or validation of the model, and incomplete data being imputed. It can be anticipated that there would be difficulties in controlling the flow of data to and from such a black box, and in monitoring what was happening inside it.

Overall, it was felt that the NN system built by NTL for the second trial worked well. However, the results of the operational evaluation were unclear on whether this system was viable. This was mainly because of doubts about whether NTL's proposed "scaled up" version of this system could cope with the volumes of census data expected in 2001.

6. STATISTICAL EVALUATION OF THE SECOND TRIAL

As mentioned earlier, a further objective of the second "Operational" trial was to carry out a rigorous statistical evaluation of the performance of the NN imputation method. For census data any imputation process must a) retain the structure of the data and b) impute plausible values. Consequently in our evaluation of the NN imputation method we examined how well marginal and joint distributions of the test data were reproduced as well as the plausibility of the actual imputations.

For comparison purposes the same data were fed through a process that mirrored the hot deck method used in the 1991 Census.

6.1 Testing Preservation of Marginal Distributions

Preservation of the marginal distribution of a categorical variable with $p + 1$ categories, the last being a reference category, can be evaluated by using the following statistic to test for marginal homogeneity in a $p \times p$ table:

$$W = \left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)'\right]\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)(\hat{y}_i - y_i)'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)\right]$$

where

y_i = p-vector indicating which of p categories the i-th individual falls into ($y_{ik} = 1$ indicates the i-th individual falls into category k, otherwise $y_{ik} = 0$).

$\hat{y}_i$ = the imputed value of y_i.

This is Stuart's (1955) extension to the case of a general square table of McNemar's test statistic (without a continuity correction) for marginal homogeneity in a 2 × 2 table. It is a Wald statistic, and, if the hypothesis of marginal homogeneity is true, should have a (large n) χ_p^2 distribution.

For continuous variables, such as age, the approach was to treat them as categorical (collapsing the tail categories) and use the above approach.

An alternative criterion is accuracy of imputation. As census output is largely tabular this was felt to be of lesser concern. Accuracy can be checked by noting the proportion of the imputations that are correct.

6.2 Evaluating Plausibility

Plausibility here means that imputed values should obey the pre-defined edit rules that are implicit in the complete data. For example in the 1991 data:

- When TYPE OF ACCOMMODATION is a ONE ROOMED FLAT, ROOMS must equal 1.
- When AGE is under 16, MARITAL STATUS must be SINGLE.

6.3 Empirical Results on Marginal Distributions

Each of the six imputed variables were examined to see how well the NN and hot deck methods preserved marginal distributions for the whole dataset and by area. The values of the Wald statistic W underpinning these results are summarized in Table 2.

The W-values in Table 2 show that for the household variables: BUILDING TYPE and ROOMS, the hot-deck method preserves the marginal distribution far better than the NN method. For the person variables there is little difference between the two imputation methods for COUNTRY OF BIRTH and PRIMARY ACTIVITY LAST WEEK while for MARITAL STATUS the NN outperforms the hot deck. However for AGE, which is a key variable, the NN performs particularly poorly.

Examination of the cross tabulations of the actual and imputed values for each variable shows that the NN is not as good as the hot deck at predicting true values.

On this evidence we concluded that hot deck performed better than the NN, although it still did not preserve the marginal distributions very well.

Table 2

W statistic values (marginal distributions)

	NEURAL NETWORK	HOT DECK
BUILDING TYPE	31.91*	7.3
NUMBER OF ROOMS	640.6*	47.6*
AGE	206.7*	41.3*
MARITAL STATUS	41.2*	89.8*
PRIMARY ACTIVITY LAST WEEK	391.8*	395.7*
COUNTRY OF BIRTH	49.7*	46.7*

* indicates significance at the 5% level

6.4 Empirical Results on Plausibility

By definition, the test data set used in the second trial was consistent with census edit rules (since it was based on "clean" 1991 data). Consequently, it was felt that if the NN can detect complex relationships in these data then it could reasonably be expected to also detect that certain combinations of values are not possible. However examination of cross tabulations of variables showed that this did not happen. The NN imputed a large number of implausible values. In contrast, the hot deck method is designed to provide imputations consistent with the edit rules and, from our analysis of the test data, seems to have achieved this aim.

To illustrate, Table 3 shows the neural network is likely to impute 2 or more rooms when the type of accommodation shows there should only be one room.

Although not shown, the NN does however appear to recognize that those under 16 years old have to be single (according to the edit rules), with substantially fewer cases incorrectly imputed.

6.5 Discussion

Analysis of the versions of the hot deck and the NN that were compared in the second trial, clearly show that the hot deck outperformed the NN both with respect to maintaining marginal distributions as well as producing "internally consistent" data. This lack of consistency of NN is particularly worrying because this method is supposed to be capable of recognizing complex interdependencies present in the data. The test data were fully consistent with the edit rules but the NN did not seem to be able to recognize these fairly obvious relationships.

On the other hand, NN did give fairly consistent imputations of marital status with age and this does point to an ability to recognize relationships in the data , thus the inconsistencies observed are surprising.

7. CONCLUSION

The results of the evaluation do not give any clear indication that the NN approach is not viable operationally. The main problem lies with whether the volume of data examined in the trial was enough to show the proposed system could cope with Census data. Any procurement exercise would have to involve a demonstration that NN can cope with a larger data set.

On the other hand NN is a 'black box' system and the results of the statistical analysis in the previous section clearly show that the NN developed for the trial was a poor performer with respect to imputation.

While ideally ONS could continue to work with NTL to build on some of the successes of the trial, while optimizing the NN with respect to the statistical quality measures used in the trial, time is now a major consideration. If the ONS is to adopt a neural approach for the 2001 census, then a NN procurement exercise has to commence immediately. There is no guarantee that the NN generated through such a process would perform any better than the one evaluated in the trial.

Table 3

Neural network imputation of rooms: Shaded areas indicate inconsistencies

		Part of converted/shared accommodation				Not part of converted/shared accommodation	Total
		1 room, self contained	1 room not self contained	2+ rooms, self contained	2+ rooms not self contained		
ROOMS	1	52	142	225	33	655	1,107
	2	29	67	248	37	1,038	1,419
	3	23	60	349	35	1,858	2,325
	4	22	33	367	32	2,393	2,847
	5+	28	52	525	53	7,140	7,798
Total		154	354	1,714	190	13,084	15,496

Consequently, it has been decided to devote ONS resources for a Census 2001 imputation system into improving the hot deck method, particularly through the introduction of a nearest neighbor/donor imputation front end. A prototype donor imputation system currently under development shows considerable improvement on the 1991 hot deck system – even maintaining marginal distributions.

Clearly, while carrying out this research we have gained considerable knowledge about the issues surrounding imputation. In particular, we still believe there is merit in the NN approach, however our evaluation, by default, focused on a particular NN implementation with poor "operating" characteristics as far as imputing missing census data is concerned. An issue that has come out of this work is that of ONS commissioning research with a private company, where there may conflict between commercial confidentiality and our desire to communicate our methods to census users. As a result of these considerations and results from similar work ONS will be carrying out further research jointly with the University of Southampton to develop an "open" NN imputation system for future censuses, based on publicly available and/or project developed software.

REFERENCES

Bishop, C.M. *Neural networks for pattern recognition*. Clarendon Press: Oxford.

Cheng, B., and Titterington, D.M. (1994). Neural networks: A review from a statistical perspective. *Statistical Science*, 9, 2-54.

Neural Technologies Limited (1996). *Final Report from Phase 1 of the Neural Imputation Trial.*

Neural Technologies Limited (1997). *Final Report from Phase 2 of the Neural Imputation Trial.*

Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics*, 12, 385-401.

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

1996 CANADIAN CENSUS DEMOGRAPHIC VARIABLES IMPUTATION

M. Bankier, A.-M. Houle and M. Luc[1]

ABSTRACT

A New Imputation Methodology (NIM) was used in the 1996 Canadian Census to carry out hot deck Edit and Imputation (E&I) for the demographic variables. The NIM allows, for the first time, the simultaneous imputation of qualitative and numeric variables. New concepts are used by the NIM to increase the quality of imputation actions. One major innovation is the identification of potential couples prior to the hot deck imputation so that special edit rules can be applied. In some cases the identification of couples involves an application of deterministic imputation, where the "deterministic action" is to blank out rather than to impute variables. This combination of deterministic and hot deck imputation is applicable to a wide range of imputation problems. The objectives and the methodology of the NIM will be briefly outlined. The identification of potential couples will be presented and its impact will be illustrated by the imputation actions for specific households.

KEY WORDS: Minimum change hot deck imputation; Inconsistent response; Couple edit rules.

1. INTRODUCTION

Among the questions asked of every Canadian on Census Day are the questions related to the demographic variables age, sex, marital status, common-law status and relationship to Person 1. The responses given to these questions are examined simultaneously for all persons in a household to identify missing and inconsistent responses. A New Imputation Methodology (NIM) was used in the 1996 Canadian Census to carry out Edit and Imputation (E&I) for these variables. This methodology allows, for the first time, minimum change imputation of numeric and qualitative variables simultaneously for large E&I problems.

Section 2 describes the basic steps of the E&I process of the demographic variables. The changes brought to the imputation system of these variables since the 1991 Census are described in Section 3. In Section 4, common response errors involving possible couples present in a household are described and illustrated by examples. Section 5 presents the solution developed to handle more effectively these frequent situations. This solution is a two-step process. The first step is a prederive module called REORDER 7, which is a major innovation in the editing of couples compared to previous censuses. This module is followed by the Edit and Imputation module. Section 6 illustrates the impact of REORDER7 and Section 7 provides some concluding remarks. More information on the NIM is available in Bankier, Luc, Nadeau and Newcombe (1996).

2. BASIC STEPS FOR EDIT AND IMPUTATION

The basic steps in the Edit and Imputation process of the demographic variables are the following:

(1) The set of edit rules which describes the invalid and inconsistent responses is determined. These rules are based on traditional assumptions that should reflect the characteristics of the Canadian population. The set of edit rules is the tool used to recognize households with missing and inconsistent responses for which imputation is required.

(2) The editing process separates the households in two groups: the households which do not match any of the conflict edit rules, and those which match at least one conflict rule. The households in the latter category are designated as "failed edit households", and the households which don't match any edit rules are identified as "passed edit households".

(3) In the imputation process, each failed edit household is corrected using the responses from a passed edit household that resembles the failed edit household.

The changes brought to the Edit and Imputation process are related to these three parts of the process. These changes are described in the next section.

3. CHANGES IN THE PROCESSING OF THE DEMOGRAPHIC VARIABLES SINCE THE 1991 CENSUS

The changes brought to the E&I process of the demographic variables are of two types. Firstly, the edit rules have been modified so as to reflect more accurately the changes in the Canadian family structures over the last few decades. One noticeable change is the importance of blended families which should be taken into consideration. Secondly, the E&I methodology itself has been modified

[1] Michael Bankier, Anne-Marie Houle and Manchi Luc, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

and NIM has replaced CANEDIT which was used since the 1976 Census. NIM has been developed to improve the methodology and to deal more effectively with the frequent response errors not well resolved by the E&I system used previously. The modifications of the edit rules and the changes in methodology are related because an important group of the new rules are part of new concepts used by the NIM. The modifications of the rules are briefly described in the next paragraphs while the following sections focus on the E&I methodology itself.

First of all, the rules comparing the ages of persons in a household have been relaxed to preserve blended families. For example, up to the 1991 Census, it was required that both parents be sufficiently older than their children. In 1996 it is only required that one parent is sufficiently older than a child. Furthermore, in 1991, numeric variables could not be used and consequently the decade of birth had to be used in the edit rules. A household failed if "The decade of birth for a son or daughter is the same or precedes the decade of birth reported for Person 1 or Person 1's spouse". In 1996, the NIM allows the use of the numeric variable age in the edit rules. Consequently, a household passes the edit rules if at least one parent is at least 15 years older than a child. Overall, the 1996 rules allow more appropriate control on the age difference between persons in a household: the use of the variable age allows much more accurate comparisons of the ages of persons, while the modification of the concepts allows the presence of blended families.

Besides the rules comparing ages, another important group of rules are the couple edit rules. These are new rules used to verify more accurately the characteristics of different types of couples, in particular the couples living in a common-law relationship which now represent an important proportion of the couples in the population. These rules will be illustrated in Section 5.1.

4. FREQUENT RESPONSE ERRORS

In the households identified as failed edit households, some common response errors are observed. First, Person 1's spouse is sometimes reported as a *son/daughter*. In the households with such an error, the difference between the age of Person 1 and the age of the "erroneous" *son/daughter* can be smaller than the accepted difference between the age of a parent and the age of a child, which causes an inconsistency problem. Another frequent situation, which is not an error but which needs attention, is when Person 1's *spouse* is not reported in position 2 on the questionnaire, even if it is specified in the questionnaire. If it is not possible to identify this person as Person 1's spouse and then make sure that the marital status responses and the common-law status responses of this person and of Person 1 are appropriate, there could be a loss of legitimate couples. Indeed, the rules that verify the characteristics of a couple formed by Person 1 and his/her spouse are applied only to the first two persons in a household so as to reduce the number of edit rules required. The household displayed in Table 1 illustrates the two problems described. In this household, Person 1's spouse is reported as a *son/daughter* and, moreover, this person is in position 5.

Table 1

Person 1's Spouse Reported as Son/daughter and not in Position 2

Relationship	Marital Status	Common-law Status	Age		CANEDIT	NIM	
Person 1	Divorced	NO	35		45		Married
Son	Single	NO	8				
Daughter	Single	NO	12	-->			
Son	Single	NO	15				
Daughter	Single	NO	36			P1's husband/ wife	Married

For this household, the minimum change imputation action is to change one variable: either the age of Person 1, the age of the last daughter or the relationship of this person. With CANEDIT and the previous age edit rules, the decade of birth of Person 1 was increased by one. On the other hand, the NIM identified the last daughter as Person 1's spouse and also changed the marital status of Person 1 and of the last person to *married*. More than the minimum number of variables was imputed by the NIM while CANEDIT imputed only one variable. In this situation, imputing the minimum number of variables is not the right decision.

Another problem is the editing of households with multiple possible couples with the same relationships to Person 1. The household presented in Table 2 is an example of this situation.

Table 2

Household with Couples with Non-unique Relationship to Person 1

Relationship	Sex	Marital Status	Common-law Status	Age
Person 1	M	Married	NO	56
Person 1's Wife	F	Married	NO	55
Son	M	Married	NO	32
Son	M	Married	NO	34
Son	M	Single	–	30
–	F	Married	NO	26
–	F	Married	NO	30
Son	M	Married	NO	27

This household presents a complex situation because there are three sons married and two married women. Consequently there are many pairs of persons that could form couples. The persons the most likely to be couples should be identified and should have appropriate responses for a couple after imputation. The large number of edit rules that would be required to handle this situation makes it necessary to have another strategy to deal with this problem. The solution developed, which is a 2-step process, is presented in the next section.

5. THE E&I SYSTEM: A 2-STEP PROCESS

The first step is a prederive module, called REORDER 7, in which potential couples are identified prior to imputation. The second step is the hot deck imputation where couple edit rules are applied to the potential couples to confirm whether these pairs are, in fact, couples.

5.1 REORDER 7

A score is assigned to each possible pair of persons in the household based on the unimputed responses to all the demographic variables. The score given reflects the likelihood for the pair of being a real couple. The pairs with the highest scores are retained where a person can belong to only one potential couple. These couples retained are identified by a person level variable, COUPLE, that is set to the same value for the two persons of a specific couple so the couple can be recognized by the NIM. In addition, for each couple retained, if the two relationships are not appropriate for a couple and if one person is related to Person 1 but the other person is not (such as *lodger* or *roommate*), this second relationship is set to blank by REORDER 7. This increases the chances for the NIM to impute an appropriate response.

The household presented in Table 3 illustrates how REORDER 7 works. In this household, the persons in positions 3 and 4 are opposite in sex, have appropriate ages for a couple but one is the daughter of Person 1 while the other one is reported as a *lodger*. These two persons were identified as a couple by REORDER 7 (as indicated by the variable COUPLE that is set to the same value (12) for the two persons) because all the variables are appropriate for a couple except the relationships. Since one relationship is related to Person 1 and the other is not, the relationship not related to Person 1 is set to blank by REORDER 7 to allow the NIM to impute an appropriate value.

Table 3
Example of a REORDER 7 Action

Relationship	Sex	Marital Status	Common-law Status	Age	COUPLE variable		REORDER 7
Person 1	M	Single	YES	53	10		
P1's C-law Partner	F	Single	YES	51	11		
Lodger	M	Single	YES	29	12	--->	Blank
Daughter	F	Single	YES	26	12		
Grandchild	M	Single	NO	7	5		

If REORDER 7 blanks out a relationship this forces the NIM to impute a value. For the household presented in Table 3, if the imputed value for the relationship is appropriate for a couple, the couple will be preserved. The other possible outcome is that a relationship not appropriate for a couple is imputed in which case the responses for the common-law status question for persons 3 and 4 have to be changed to *NO* to be consistent with the relationships to Person 1 of the two persons.

Therefore, REORDER 7 reduces the number of NIM edit rules required because the couple edit rules in the NIM are applied only to the couples identified by RECORDER 7. In addition, RECORDER 7 allows correct imputation actions to be achieved when more than the minimum number of variables have to be imputed.

The combined process of REORDER 7 and NIM can be viewed as a combination of deterministic imputation and hot deck imputation. The fact that relationships are set to blank is a form of deterministic imputation, where the "deterministic action" is to blank out rather than to impute variables. This allows more plausible imputation actions, through the imputation of more than the minimum number of variables. This combination of deterministic and hot deck imputation is applicable to a wide range of imputation problems. The common concept of these problems is the development of a strategy to determine the optimal variables to blank out so as to guide the hot deck imputation to the most plausible imputation action. Such a strategy is illustrated by a modification that could be made to the RECORDER 7 module for the 2001 Census. Since the NIM performs minimum change imputation in some sense (Section 5.2), it will tend to eliminate couples who give no indication that they are legally married or in a common-law relationship. For example, the household of Table 4 fails because Person 1 and the common-law partner of Person 1 have the response *NO* for the common-law status question. In such a case, the NIM will generally change the relationship of the second person rather than change the common-law status of the two persons, because only one variable is imputed instead of two.

Table 4
Failed Edit Household

Relationship	Marital Status	Common-law Status
Person 1	Separated	NO
Common-law Spouse of P1	Separated	NO

In this situation, where it is believed that these two persons possibly form a couple, blanking out variables would increase the chance of the couple being retained by the NIM. One possibility would be to blank out the common-law status *NO* of the second person in which case two variables have to be imputed to retain the couple or to eliminate it. In this case, if the common-law status response *YES* is imputed for Person 2, the common-law status response for Person 1 has to be changed to *YES*. On the other hand, if the response *NO* is imputed for Person 2 then the relationship of Person 2 also has to be modified. From the perspective of minimum change, the two imputation actions are equally attractive because two variables are imputed in either case. Thus the frequency with which the couple is retained will be based on the frequency with which such couples appear among the donors. This way of improving hot deck imputation by deterministically blanking out variables prior to imputation could be applied to a broader range of imputation problems.

After the identification of the potential couples in REORDER 7, the Edit and Imputation process is performed by the NIM. At the editing step, the set of edit rules is applied to all households to identify the households for which imputation is required. The couple edit rules represent an important group of these rules but they are applied only to the couples identified by REORDER 7. An example of these couple edit rules applied in the NIM are illustrated in Table 5 for the "*son/daughter-son/daughter-in-law*" couples. Edit rules similar to those presented in this table exist for pairs of persons with other relationships that could form couples (for example a *brother/sister* and a *brother/sister-in-law*).

The rules illustrated in Table 5 are generated by the NIM Edit Interface for all the combinations of two persons in the household. The quantities "#1" and "#2" represent any combination of two persons. The first proposition ensures that the rules are applied only to the couples identified by RECORDER 7.

Table 5
Between Persons Edit Rules for "Son/daughter – Son/daughter-in-law" Couples

Propositions	Rules								
	1	2	3	4	5	6	7	8	9
COUPLE #1 = COUPLE #2	Y	Y	Y	Y	Y	Y	Y	Y	Y
relationship #1 = S/D	Y	Y	Y	Y	Y	Y	Y	Y	N
relationship #2 = S/D-in-law	Y	Y	Y	Y	Y	Y	Y	N	Y
Sex #1 = sex #2	Y								
marital status #1 = married		Y	N			N			
marital status #2 = married		N	Y				N		
common-law status #1 = yes				Y	N	N		Y	
common-law status #2 = yes				N	Y		N		Y

The definition of a couple requires that the two persons of a couple be opposite in sex and that both be married or living in a common-law relationship. The set of rules illustrated in Table 5 ensures that the couples respect these conditions. If two persons with appropriate relationships for a couple, and identified as a couple by REORDER 7, match one of these edit rules, there are two possible outcomes: either the variables that caused the household to match the edit rule are changed so as to be appropriate for a couple, or the relationship of one person is changed such that the relationships are no longer appropriate for a couple. The pair will then not be considered any longer as a couple.

After the identification of the potential couples present in the households and the application of the edit rules to identify the households with missing and inconsistent responses, the last step of the process is the imputation of these failed edit households, which is discussed in the next section.

5.2 The New Imputation Methodology

A new imputation methodology, NIM, was used in the 1996 Canadian Census to carry out Edit and Imputation of the demographic variables. This method is based on the principle of minimum change while taking into consideration the plausibility of the possible imputation actions. NIM allows, for the first time, imputation of numeric and qualitative variables simultaneously for large E&I problems. At the same time it deals more effectively with the most frequent errors made by respondents.

The objectives of an automated hot deck imputation methodology should be as follows:

(a) The imputed household should closely resemble the failed edit household.

(b) The imputed data for a household should come from a single donor, if possible, rather than two or more donors. In addition, the imputed household should closely resemble that single donor.

(c) Equally good imputation actions, based on the available donors, should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population.

These objectives of an imputation methodology are achieved under the NIM by first identifying the passed edit households which are as similar as possible to the failed edit household. These households are designated by the name "nearest neighbours". For each nearest neighbour, NIM attempts to impute each combination of variables that don't match the failed edit households. One of these imputation actions which passes the edits and resembles both the failed edit household and the passed edit household is then randomly selected. The next sub-sections describe in more detail how these three procedures are realized.

5.2.1 Distance Between Failed and Passed Edit Households

The notion of similarity is based on a distance measure. It will be assumed that F households match at least one conflict edit rule, while P households don't match any edit rules. The responses for the households that failed and passed the rules are labelled respectively by $V_f = [V_{fi}], f = 1$ to F and $V_p = [V_{pi}], p = 1$ to $P, i = 1, \ldots, I$. These are $I \times 1$ vectors containing the responses for all the persons in a household, where I will vary according to the household size. Each failed edit household is compared to each passed edit household. The distance between each failed edit household V_f and each passed edit household V_p is defined as

$$D_{fp} = D(V_f, V_p) = \sum_{i=1}^{I} \omega_i D_i (V_{fi}, V_{pi})$$

The distance function $D_i(V_{fi}, V_{pi})$ can be different for each variable i. In the 1996 Census, one function was used for qualitative variables while a second function was used for the numeric variable age. For the qualitative variables, $D_i(V_{fi}, V_{pi}) = 1$ if V_{fi} is not equal to V_{pi}, and $D_i(V_{fi}, V_{pi}) = 0$ otherwise. For the numeric variable age, $D_i(V_{fi}, V_{pi}) = 1$ if $| V_{fi} - V_{pi} | \geq 6$, and $0 \leq D_i(V_{fi}, V_{pi}) < 1$ if $| V_{fi} - V_{pi} | < 6$. For each failed edit household, the D passed edit households with the smallest distances are considered as potential donors for the failed edit household. The different possible imputation actions based on these D potential donors are generated and one of them is selected to be the actual imputation action for the failed edit household.

5.2.2 Possible Imputation Actions

Each passed edit household V_p will mismatch the failed edit household V_f on one or more variables. Let I_{fp} represent the number of variables which mismatch for V_f and V_p. Since each of these I_{fp} variables can take the value of the failed edit household or the value of the passed edit household, but that at least one variable must be imputed, there are $2^{Ifp} - 1$ possible imputation actions for the failed edit household V_f using the passed edit household V_p. If the D potential donors are considered, the quantity $N_f = \sum(2^{Ifp} - 1)$ represents the total number of possible imputation actions for the failed edit household V_f. It is then necessary to have criteria to select one of the imputation actions from one of the passed edit households. This selection is based on a weighted distance between the failed edit household, the passed edit household and the imputation action.

5.2.3 Selection of an Imputation Action

For each possible imputation action which passes the edit rules, the following weighted distance is calculated:

$$D_{fpa} = \alpha D(V_f, V_a) + (1 - \alpha)\, D(V_a,\, V_p)$$

where $D(V_f, V_a)$ is the distance between the failed edit household and the imputation action (measure of the number of variables imputed), and $D(V_a, V_p)$ is the distance between the imputation action and the passed edit household (measure of the plausibility of the imputation action). The parameter α takes a value between 0 and 1. In the 1996 Census, $\alpha = 0.9$ was used. Therefore, emphasis is placed on minimizing $D(V_f, V_a)$ rather than minimizing $D(V_a, V_p)$. This weighted distance is calculated for each potential imputation action and is used to select an imputation action which respects the objectives of an hot deck imputation methodology.

Of all the imputation actions considered which pass the edit rules, only those which minimize or nearly minimize the weighted distance are retained. The imputation actions retained must satisfy $D_{fpa} \leq \gamma \min D_{fpa}$ where $\gamma \geq 1$. In the 1996 Census, γ was set equal to 1.1. These imputation actions are called "near minimum change imputation actions", while the imputation actions with $D_{fpa} = \min D_{fpa}$ are called "minimum change imputation actions". In other words, imputation actions which don't exactly minimize the weighted distance are allowed because these imputation actions might be as plausible, and sometimes more plausible, than the imputation actions which minimize the distance. In addition, the imputation actions retained are such that no subset of the variables imputed represent another imputation action which would pass the edit rules. If this was the case, one or more variables would be unnecessarily imputed, which would violate the principle of making as little change to the data as possible. These imputation actions retained are called potential imputation actions.

A size measure defined as $R_{fpa} = (\min D_{fpa} / D_{fpa})^t$ is calculated for each potential imputation action. The parameter t is there to give more or less weight to the minimum distance imputation actions as opposed to the near minimum change imputation actions. This parameter was set to 1 in the 1996 Census. One of the potential imputation actions is randomly selected, with probability proportional to R_{fpa}, to be the actual imputation action for V_f. This completes the selection of an imputation action with the NIM.

The impact of the module REORDER 7 and of the use of the couple edit rules is illustrated in the next section with the sample of 1/5 private households in Canada who received the long form questionnaire.

6. ILLUSTRATION OF THE IMPACT OF RECORDER 7 AND OF THE COUPLE EDIT RULES

To study the effect of the identification of couples prior to imputation and of the application of couple edit rules to the couples identified, the "*son/daughter – son/daughter's partner*" couples were studied for a sample of private households. The study was restricted to this category of couples because they are one of the most frequent type of couples.

There are 22,350 couples in the category "*son/daughter – son/daughter's partner*" couples identified by RECORDER 7 in the sample considered. These couples are part of both passed edit households and failed edit households. For 97% of these couples, the two relationships are present after REORDER 7. Therefore one relationship is missing in only 3% of these couples. Of the 22,350 couples identified by REORDER 7, 83% were retained by imputation. The fact that a couple is preserved or not by imputation is related to the responses to the other variables. For 98.6% of the 18,522 couples retained, both persons are reported as *married* or both have common-law status *YES* before imputation. In addition, for 98% of the couples retained, both persons are older than 15 years old before imputation. The responses to these variables thus suggest that the persons form a couple. On the other hand, for 90% of the couples not retained by the NIM, both persons are not reported as *married* and both have common-law status *NO* or missing before imputation. Finally, for 58% of the couples not retained, at least one person is less than 15 years old before imputation. An example of a couple retained by the NIM is given in Table 6. This household was previously used in Section 5.1 to illustrate the impact of REORDER 7. The relationship *lodger* in position 3 was set to blank because the responses for the persons 3 and 4 suggest that these two persons form a couple. Blanking out the relationship allows the NIM to consider these two persons as a potential couple, to apply edit rules to verify their status as a couple and to correct responses if necessary. In this case, NIM imputed the relationship *son-in-law* which is plausible considering the structure of the household.

Table 6
Example of a Couple Created by REORDER 7

Relationship	Sex	Marital Status	Common-law Status	Age		REORDER 7		NIM
Person 1	M	Single	YES	53				
P1's CLP	F	Single	YES	51				
Lodger	M	Single	YES	29	-->	blank	-->	son-in-law
Daughter	F	Single	YES	26				
Grandchild	M	Single	NO	7				

To evaluate the relative importance of the identification of couples prior to imputation it is also important to examine the couples present in the households after the E&I process. In the sample of households considered, there are 18,756 "*son/daughter – son/daughter's partner*" couples after imputation. Of these couples, 99% were identified by RECORDER 7. Most of these 18,756 couples after imputation (86%) didn't have any variable changed. For 14% of the couples present after imputation at least one variable was imputed, either because of non-response or because of inconsistencies. The combination of RECORDER 7 and the NIM, consequently, had some impact on about 14% of the "*son/daughter – son/daughter's partner*" couples in this sample.

The additional feature of the NIM of identifying the potential couples and verifying their characteristics requires a considerable increase in the number of edit rules. The following table presents the number of edit rules required to process the households of each size. The last column presents the standardized time to process a household of a specific size in terms of the time to process a household of size 1.

Table 7
NIM cost as Household Size Increases

Household Size	Number of Edit Rules	Standardized Time in terms of Time for 1-Person Hhld.
1	9	100
2	49	129
3	307	230
4	787	459
5	1,494	566
6	2,435	1005
7	3,616	1182
8	5,043	941

The processing time does not increase proportionally to the number of edit rules. In particular, the time does not increase between the 7-person households and the 8-person households because of the reduction in the number of possible donor households. The fact that the time does not increase proportionally to the number of edit rules is a considerable improvement compared to the method used up to the 1991 Census. With this previous method, the processing time was increasing exponentially as a function of the number of edit rules. Therefore, in addition to respecting the objectives of an hot deck imputation methodology, NIM verifies more thoroughly the demographic characteristics of the persons and is also much more efficient from an operational viewpoint.

7. CONCLUDING REMARKS

One of the major innovation of the NIM is the identification of couples followed by the minimum change imputation of the demographic variables. This process was computationally feasible and effective and contributed to increase the data quality.

For the 2001 Census, improvements will be made to the NIM methodology for the demographic variables. For example, instead of identifying potential couples prior to imputation, as the actual REORDER 7 module does, the possibility of identifying potential families will be evaluated. This would allow the application of more detailed edit rules to verify the characteristics of families. In addition, the NIM will be generalized for the 2001 Census so it can process a wider selection of variables.

REFERENCES

Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1996). Imputing numeric and qualitative census variables simultaneously. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1996.

SESSION I-9

Data Quality and Improving Data Processing

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

DATA CAPTURE AND CODING FOR THE 2001 GB CENSUS

R. Massingham[1]

ABSTRACT

The Office for National Statistics (ONS) and the General Register Office for Scotland (GRO(S)), co-operating closely with the Northern Ireland Statistical Agency (NISRA), are carrying out the first major Census Test in the United Kingdom since 1991 to assess the response to new questions and evaluate the results of a number of new and previously untried systems. One of the primary activities of the 1997 Census Test is to trial a prototype data capture and coding system. The key elements being tried for the first time in GB include automatic scanning and recognition of forms completed by the public, capturing and automatically coding the data with interactive coding query resolution. This has involved in-house development as well as the integration of jointly developed software packages. The Test processing operation is controlled by a processing control system working at census form level. The approach for 1997 has been designed after carrying out extensive research and a series of mini-trials in the past 2 years. The 1997 Census Test will help to determine the processing strategy for the 2001 Census, and to decide the approach for the acquisition of a fully integrated processing system, with the opportunity to refine the chosen system after a major dress rehearsal in 1999.

KEY WORDS: Census test; Data capture; Coding; Outsourcing.

1. TOWARDS THE 2001 CENSUS

A number of important lessons were learnt from the processing of the last census held in Great Britain. The processing operation was considered a qualified success. The key dates were met but some re-processing had to be carried out because of a data problem. The systems required to capture, code, and edit the data were complex and very resource intensive. The serial batch processing flow was not ideal and sometimes delayed processing until batches were complete. The demands for more timely and consistent 2001 output of the best possible quality and to produce this within the same order of costs as those for the 1991 Census have set a major challenge for the Census Offices of Great Britain, the Office for National Statistics and the General Register Office for Scotland.

The almost certain use in the 2001 Census of data capture technology involving scanning and recognition and automatic coding will mark another significant change in census processing methodology in Great Britain. Previous important landmarks included the use of electro-mechanical counters and sorters in the 1911 Census; the use of a mainframe computer in 1961; and processing the data by keying directly to disc in 1981. Until now the data "vehicle" has been the completed paper census form. For the first time in census-taking history in Great Britain, the first census of the new Millennium will involve the removal of the paper form, once scanned, from processing. This important change will lead to enormous benefits and at the same time present new challenges to processing methodology. Advances in appropriate technology and rising staff costs underpin the need to explore new and innovative ways of delivering reliable products to an increasingly demanding customer base.

2. 2001 CENSUS

Strategic Aims

Before looking more closely at the plans and options for 2001 Census processing, it is necessary to set them in context with the high-level strategic aims, relevant planning assumptions and the research which has been conducted since the 1991 Census.

There are 4 aims:

- to ensure that the question context is appropriate; to meet the demonstrated requirements, taking account of considerations of value for money;
- to deliver census products and services to meet legal obligations and customers' needs within stated quality standards and to a pre-defined timetable;
- to ensure that all aspects of the Census operation, and the dissemination of results, are acceptable to the public and comply with Data Protection law;
- to demonstrate that the Census represents value for money.

These aims demand a well-planned and executed census operation, incorporating as seamless a flow as possible from data collection through to processing for the production of a clean and consistent database from which outputs can be produced.

Planning Assumptions

A number of planning assumptions, relating to the strategic aims, have been produced based on 1991 Census experience and updated with current research and testing experience. The ones most affecting the approach to 2001 processing methodology are:

[1] Rod Massingham, Office for National Statistics, Titchfield, PO15 5RR England; e-mail: rod.massingham@ons.gov.uk.

- There will be 10% more households to enumerate than in 1991:
  Brought about principally by the increasing number of one person households; there will be nearly 2 million more forms to process.
- Information collected in 2001 will be processed at 100%:
  This will be a first-time event; although collected from every household, some census data, have been processed at 10%. These cover the occupation and industry topics which are labour-intensive to code. In 2001, processing these topics at the 100% level presents one of the biggest challenges to the Census Offices. Clearly, it is uneconomic to code ten times the amount manually, but even automatic coding engines leave, according to experience so far, some 30-40% to be resolved interactively. This still represents, a four-fold increase in the task over 1991. The cost implications are obvious.
- As the Census is a high-risk one-off operation, with no second chance to recover, the whole system including the processing systems will be dress rehearsed:
  There will be a 1999 Dress Rehearsal, in which a full test of the chosen processing system will be conducted, after which the system will be refined and scaled up for 2001.
- Technological advances offer new opportunities which will be evaluated and built into the programme where proven and cost effective:
  Scanning, recognition and automatic coding techniques are being tested in the 1997 Census Test in UK. This evaluation will help to develop the 2001 processing strategy.
- Private finance initiatives, outsourcing and partnership arrangements will be used where proven to be cost-effective, within the constraint of confidentiality:
  These will continue to be developed, building on previous practice; key decisions will be made soon regarding the extent to which processing systems can be contracted out in the context of the relatively new, at least to the UK Census Offices, and rapidly developing technologies.

3. RESEARCH SINCE 1991

Background

The 3 UK Census Offices conducted a policy evaluation and reappraisal of customer requirements for, and alternative methods of, providing census-type data for households and persons for the period 1996-2016. The outcome was a preference for a traditional census carried out by a central agency and funded by government. Data similar to 1991 was required for the foreseeable future and outputs should be timely and reliable with key counts available within 6 months and pre-defined outputs within one year.

In a review of the 1991 Census, the National Audit Office commented that development work ought to have started earlier. As census-taking was becoming more difficult because of social and behavioural changes, new and innovative methods needed to be explored. Consequently a census development programme was started to review and evaluate 1991 performance and to investigate enhancements to improve the cost effectiveness of the 2001 Census. Consideration was also given to work which might be contracted to the private sector to achieve efficiency savings but without risk to the Census. One of the potential activities for outsourcing was data capture and coding.

Studies were subsequently conducted into data capture and coding methods, and in-house trials between 1995-1996 included imaging and its management, recognition, and keying from image, and coding software packages.

Main Aims

The processing research programme included the following key aims with appropriate approaches where applicable.

Streamline the capture and coding process with a view to reducing staffing costs and number of processes, and therefore accommodation required.

- feasibility of image manipulation to optimise capture and coding functions at person and form level;
- feasibility of using OMR to reduce clerical capture of tick boxes; and using OCR to enhance capture of text, both for numeric questions and for postcodes.

Improve quality in terms of accuracy and consistency of captured and coded data.

- compare quality and accuracy against a key and code from paper benchmark;
- feasibility of automatic coding for different topics.

Improve the speed of processing of data over 1991 so that outputs can be delivered to a reliable timetable.

- assess impact upon working practices – staffing profiles, ergonomic requirements;
- determine approach to keying and coding – whole job, specialism, and combinations of questions.

Enable the trade-off between reduced costs and processing all data at 100% level.

Summary of Key Results

Imaging was considered a viable and practicable alternative to the paper form in terms of management of processing input data and the flow of work through a system. OMR and OCR also achieved good recognition rates of ticks, numerics and presence of text. The coding trials revealed that automatic coding was also viable and that for certain topics little or no query resolution would be required thus helping to reduce costs.

This outcome helped to underpin the strategy of being able to develop a specification for work to be carried out for the 1999 Dress Rehearsal and 2001 Census. Testing a prototype processing system for the 1997 Census Test would help to confirm the research, as well as building necessary in-house skills and experience.

4. THE 1997 CENSUS TEST

This was the first full scale test to be held since 1991 and covered 100,000 households in the UK. Processing and evaluation of data from the 60,000 or so responding households (the Test was voluntary) are still underway. Evaluation of scanning and recognition will be completed shortly as will some lessons from the coding operation – these will be fed into the strategic plans now being assembled.

The key drivers of the Test were enumeration methodology, form style, and form content which are reflected in the aims.

Main Aims

- *to try out new and revised enumeration procedures* for improving coverage, quality and cost-effectiveness. These included testing postal methods of delivery and collection of census forms as well as the conventional method of delivery and collection by an enumerator;
- *to make the census form easier for people to complete*. Two different types of census form (matrix style as in 1991, and pages per person format) with new and revised questions, were tested to gauge public reaction to the layout and question wording;
- *to test the acceptability of a question on income*. Users had stated this to be their top priority for a new census question;
- *to improve the census field operation* by using Geographic Information Systems software and digital maps and boundaries for planning enumeration areas, and for supplying enumerators with a smaller scale map supplemented by an address list;
- *to improve the efficiency and effectiveness of recording the data from the census forms* using document scanning, optical mark reading and optical character recognition techniques, and automatic and computer-assisted coding of write-in responses.

Data Capture and Coding

Overview

All the completed census forms in GB are being processed by GRO(S) on a conventional key to disc basis, and the data capture is being used for the full Test evaluation. No coding of write-in answers is being carried out as this is being evaluated as part of the automatic data capture and coding system.

A prototype capture and coding system developed for the 1997 Census Test is processing 30,000 forms, about one half of the responding households, and including a small sample of forms from Northern Ireland.

The enumeration districts (collection units) selected for scanning are representative of each GB geographical area type in the Test (there are 5), and of form types (matrix and pages per person) with question variants (including the income question). There are 14 different types of forms being handled in the system.

The system approach was to scan the forms on a remote scanner in NW England, input via digital audio tapes into a capture/recognition system jointly developed with an external developer and linked to an automatic coding system using a variety of in-house and externally developed software, with query resolution being handled interactively. A system overview diagram is attached.

System Description

Logistics and form preparation

Boxes of completed forms – nearly 500 districts – were logged in and separated out for scanning. The forms were guillotined in-house (to remove staples) prior to shipment in secure transport to the scanning centre and GRO(S) Edinburgh. The forms are being stored for further Test evaluation.

Scanning

Pages of forms, uniquely serial numbered, were fed in batches into a high-volume scanner. Output in Group 4 TIFF format on DAT was transported securely to the capture system on a daily basis. This was completed at the end of September 1997. Identification numbers on the front page of each form were captured by OCR and using the serial number on each form, the "form family" was kept intact.

Rejections were high initially because of inconsistent printing problems; further tuning of the scanner overcame some, but not all, of them.

Capture

The system comprises off-the-shelf Pentium PCs with 17" monitors, operating as servers, keying/verification stations, scanning and recognition, exception stations, with an optical disk server with jukebox. Software has been developed for scanning, recognition, exception and verification by a UK based company (US parent) as well as using third party optical system software. The contractor also provided commissioning, joint development, training and maintenance/problem resolution services.

OCR is being used to capture alpha/numerics for five questions. Unrecognised characters are being keyed from image. OMR is being used to capture ticks and detect the presence of text for keying. Edit rules are being applied to resolve multi-ticked responses, and indicators are set for use by a data quality management system.

Postcodes, used in UK addresses, pose a particular capture problem because of their alpha/numeric format. The three methods being tested are (a) using OCR for address 1 year ago – being key verified as a check on OCR; (b) using OMR for workplace address where the postcode can be keyed from the image; (c) the postcode of the enumeration address is being generated from a geographic database by matching the form identity on the front of the form.

The data is reformatted into household and person records for the coding processes.

Coding

There are 3 groups: simple, address, and complex. Simple questions, country of birth and ethnic origin, are being automatically coded using in-house developed software and classifications. Address questions, already mentioned above, are being coded to postcode level using a proprietary automatic address coding software package. The two complex questions, industry and occupation, are being automatically coded using software developed in Australia, and tuned jointly to British classifications. Query resolution, via image, will be using a variety of sources, including a business register for industry. Unresolved queries will be passed to experts for further resolution.

Control

Tracking the progress of household and person records is controlled by a processing control system. This links to a forms control database from which summary management information about progress is provided. The control system prioritises work, for example, by accelerating processing of a particular geographical area or question types. The coding supervisor is then able to re-allocate work accordingly. Progress on questions passing through the system and integrity checks on person counts are also registered.

5. 1997 TEST – EMERGING ISSUES

Processing operations and evaluation of the Test have yet to be completed but issues for the processing development and operation ahead are emerging.

Imaging

Images are an acceptable presentation vehicle for keying and the new working environment, without the paper form (once scanned), and will require different approaches in terms of system flow and management and ergonomics.

Forms Disposal

The question of the possible disposal of the completed census forms arises. The national Public Records Office agreed, subject to assurances on quality, that microfilm was acceptable as the principal record and that paper forms could be destroyed. Images may also be acceptable as archival media, but the longer-term issue of image technology being available a century from now adds to the problem. Microfilming at the scanning stage is an option.

Data Needs, Form Design & Technology

The relationship between the design and content of the form and therefore public acceptability; its constituents, and the scanning requirements, importantly the choice of scanner, are essential ingredients to the success of the capture process. The processes involved in form design must be considered much earlier in the planning process than in previous censuses. Registration marks, drop-out colours, ink and paper quality, and serial numbering are some of the considerations. Coupled with this is the need to balance customer data requirements and expectations with the realistic system delivery levels in terms of quality and accuracy.

Partnerships in Acquisition and Development

Assumptions are being made on the options for which systems can be contracted out. For example, data capture (*i.e.*, scanning + recognition) could either be acquired and developed in-house, or operated in-house with system development being contracted out, or contract out the whole system. The same mix could apply to coding but the complex coding software could be jointly developed (as in 1997). The Test has provided the experience and developed skills necessary to assess and manage contracted options. The additional processes of edit and imputation could follow a similar pattern to coding. Key questions are how much and in what form should the various systems be out sourced and the management of risk in terms of delivery, census integrity and confidentiality.

6. CONTINUING RESEARCH

Further research is being conducted into scanning of the 1997 Test form, and the programme of small scale testing is continuing to try out different versions of questions, including occupation and industry. A number of concept forms, both in the matrix and pages per person styles are being developed in conjunction with users' needs. The present 32 page Test form could be reduced to around 16 pages, but even this smaller form will mean the processing of almost half a billion images in 2001.

7. OTHER PROCESSING ISSUES

Improving Technology, Software, and Techniques

Continuing development in the data capture industry will benefit the UK Census Offices. Scanners are becoming faster and more reliable and can be sophisticated in front-end processing facilities. As with all technology, a balance has to be struck between cost, risk and quality. Alternative keying methods (*e.g.*, carpet keying, whole form, or snippet) will need to be carefully assessed.

Processing Location

The option of processing at one or more sites is being addressed. Procurement of government sites for example is becoming more difficult but the size needed, compared with 1991, will be smaller. Private sector provision is also being considered. Much will depend on the route taken on outsourcing.

Distributed Processing

The distribution of the system demands careful appraisal of options because of a number of issues. These include the extent of distribution, control of processes, support – both internal and external – back-up and recovery, communications, network traffic efficiency and effectiveness, movement of electronic files and database management. Database design can be more complicated – what data client PCs and servers hold and which processes are run on each.

Free-Flow Processing

The 1991 batch processing method was slow. A more flexible approach may be required and this can be achieved by scanning forms in no pre-determined order, the key objective being to push the 25 million forms into the system as fast as possible – postal collection and the expected late returns add to the need for speedier and more flexible input for coding purposes.

Job Approach

The impact of the image environment on how interactive processing is conducted is being evaluated in the 1997 Test. Keying from image is being monitored – the "heads-up" posture and no paper for example is an acceptable environment and is preferred by traditional keyers. The issue of how coding is handled is complex – whether coders deal with the whole form (*i.e.*, coding any data presented) or by topic/question only – will need to be reviewed from the Test's findings together with other agencies' approaches. Each method impacts on the flow and control requirements by either "streaming" to topic experts or merely apportioning workloads. Quality is a key issue.

System Design and Editing

The overall processing system will take account of editing and imputation requirements, impact of the choice of the new database and development tool for ONS. The stage at which and in what form the main editing processes take place in processing needs to be considered as part of the design. An important issue is whether a single edit is practical for the new 100% coding approach, bearing in mind the timing of coding processes, or whether a two stage edit is better. The quality of the chosen OCR engines is key to the degree of calibration for the recognition processes and the impact of that on data and editing. Rules for multi-ticking, range checks, and cross question consistency must be built into the edit process. The timing of imputation, adding any missing persons and no contacts, bears on the processing design. Whichever route is taken on acquisition, the system needs ideally to be able to interface with the new ONS database, using a Census-specific database may be the better option. Effects of the One Number Census (*i.e.*, adding missing households and accessing administrative records) on earlier processes will also need to be considered.

8. 2001 PROCESSING STRATEGY AND OPTIONS

ONS is considering options for acquiring the data capture and coding system. In-house expertise in procurement is currently being assessed and linking it with various options for the system. Smaller teams with contract management, technical and financial skills with a robust quality control approach would be required. Information about processing approaches from other statistical agencies is also being reviewed. The lessons leamt so far from the research and testing programme give indications of what might be the best approach for 2001, but the best way forward is to involve potential contractors closely in helping first to identify the options for acquiring services.

The system currently envisaged is a fully integrated and automatically controlled one in which data capture and coding are seamless. This is in part the 1997 model, but more work is required to move to a 2001 system. This raises questions of whether the whole system could be contracted out. Experience suggests that contracting out scanning and data capture only may be the more practical solution with the coding being controlled by the census organisation, using jointly developed software.

There are variants: scanning only could be out sourced with little difficulty – the 1997 Test used a remote scanner. The scanning provider would receive the 25 million or so completed census forms either at one or more of their own centres or at a single, large census processing office or at 4 (say) Census separate offices. The latter has advantages of control, in terms of logistics and security – actual and apparent (to the public) – and proximity to the remaining processing systems, depending on distribution of the system architecture. Staff levels, training programmes, and accommodation type are all important considerations whichever acquisition route is preferred.

Capture and recognition entirely could be out sourced – this would be a change to the 1997 model. Experience gained from the joint development approach would be needed, however, to understand the technical intricacies, and to ensure that proposals were robust. Separate procurements of scanning and capture and recognition is less attractive.

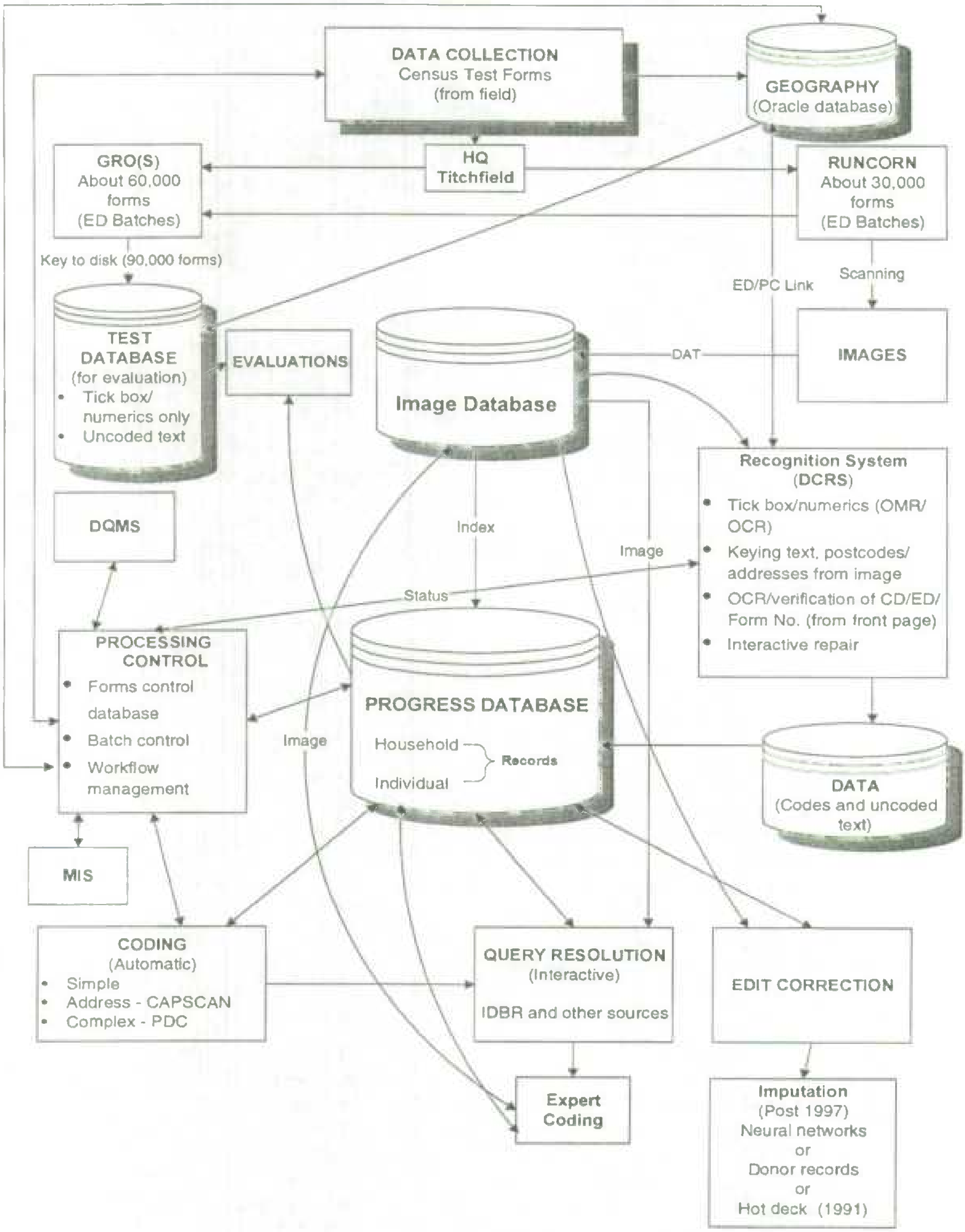
Coding presents different problems because of the expertise required not only to build the appropriate indexes and other references but also to tune software to often complex classifications. The resolution of responses not successfully coded automatically demands in-house expertise. However, two important aspects influence that thinking. Firstly, the level of accuracy and quality which would meet customers' needs and secondly, the increasing dexterity of automatic coding engines. Coding software can operate to varying levels of accuracy depending on the comprehensiveness of indexes and with techniques like fuzzy matching.

9. CONCLUSION

Post 1991 Census research, the early findings from the 1997 Census Test, and exchanges with other international agencies has confirmed the Census Offices' approach for the 2001 Census. Much information is still being gathered which is feeding into the strategic plans. There are no firm recommendations yet and options remain open. There is much to support contracting out to the private sector, but caution is also needed as the census is vital - failure is not an option. Exhaustive testing of systems, tight control on development and costs, and robust appraisal are all required if strategic aims are to be met.

The Census Offices are embarking on a challenging and interesting path for processing and will need to be firmly in touch with national and international developments. Acquiring the best available expertise and building on the offices' skills are pre-requisites to deliver the most cost effective processing system possible.

SCOPE OF 1997 CENSUS TEST PROCESSING SYSTEM OVERVIEW



Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

TOWARDS INTEGRATED BUSINESS SURVEY PROCESSING

F. van de Pol, A. Buijs, G. van der Horst and T. de Waal[1]

ABSTRACT

Statistics Netherlands is in a process of transition in business survey data processing. Firstly, there is a shift from paper form data entry to electronic data interchange (EDI). Secondly, administrative data files are increasingly used, either directly for tabulation or for matching and mass imputation. Thirdly, budget constraints call for more efficient data processing methods. To cope with these developments, 1) several years ago computer-assisted data entry (with BLAISE) was introduced, 2) an 'OK index' was devised to trace the 'critical stream', which needs extensive micro-editing, 3) GEIS-like automatic editing software, called CHERRYPI, was developed for the 'non-critical' stream, 4) pre-imputation for late response and nonresponse is used, especially in the critical stream, and 5) a tool for graphical macro-editing is developed, called MACROVIEW. The present paper highlights the interrelationships between these methods and also gives details on the latest developments: parallel computation in CHERRYPI and the functionality of MACROVIEW.

KEY WORDS: Automatic editing; Macro-editing; Edit strategy; Parallel computing; Graphical editing.

1. INTRODUCTION

A few decades ago data editing meant manual correction of paper forms. Nowadays data editing takes place during or after data entry in the computer. The present standard way of processing business survey records at Statistics Netherlands is to micro-edit every incoming form during data entry with a Blaise CADE (Computer Assisted Data Entry) application. Notwithstanding the benefits of CADE, editing all incoming forms extensively is very labour intensive. An extreme example is data processing for Annual Production Surveys, where editing one form takes several hours on the average. Moreover administrative files, like tax files, are increasingly used and these files are simply too large for extensive manual micro-editing. For these reasons we have started a project to devise methods which require less manual corrections.

In the literature several reviews of data editing methods are given (Pierzchala 1990; Granquist 1994, 1995; Granquist and Kovar 1997; Bethlehem and Van de Pol 1997). What we would like to focus on in this introduction is the sequence in which these methods are to be applied. Let us first make the distinction between micro-editing and macro-editing. Micro-editing is to detect errors on the level of records and macro-editing is to detect errors on the level of aggregates (publication figures) or distributions. Correction of errors is always on the micro level of records.

Micro-editing may reveal more errors than macro-editing, but macro-editing will trace bigger, more influential errors. Micro-editing and macro-editing are complementary. Errors that are apparent from one point of view may not be apparent from the other.

Macro-editing is generally viewed as a final check just before publication. However, when the incoming raw data have few errors, one may skip micro-editing and rely fully on macro-editing. The argument for this choice would be that errors that one misses from the macro point of view do not have to be corrected to obtain good publication figures. One counter argument to maintain some sort of micro editing is that classification variables such as branch of industry and firm size class should at least have valid values. A more principal argument against relying fully on macro-editing is that taking publication figures as point of departure for tracing errors may prevent publication of unexpected, but true changes in trend. Outliers in one direction may be removed until outliers in the opposite direction cancel out the unexpected trend. This argument is increasingly often overruled by the dictate of budget cuts, though. A final argument for maintaining some sort of micro-editing is that this is the only way to make sure that all so called 'hard errors' are eliminated. (A hard error occurs for instance when detail figures in a form do not add up to their reported total (Bethlehem and Van de Pol 1995).) This latter argument reflects a bookkeeping mentality which is not only opposed to macro-editing, but to all statistical data editing methods.

When the incoming raw data contain many errors, that is when almost every record needs correction, systematic micro-editing is more efficient. In that case the extra effort to trace erroneous records from a macro point of view should be postponed until the dataset has a reasonable good quality due to micro-editing. Macro-editing cannot be missed, because it reveals errors that will go unnoticed with micro-editing.

We choose to do micro-editing before macro-editing. Especially automatic micro-editing should take place before macro-editing, because making a correction automatically is preferable to doing it by hand, provided that the automatic correction is an improvement. With some surveys, automatic correction is only partially successful. Then

[1] Frank van de Pol, Statistics Netherlands, Department of Statistical Methods, P.O. Box 4000, 2273 TX Voorburg, The Netherlands, e-mail: fpol@cbs.nl. The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

computer assisted manual correction (CAMAC) should also be applied, but only for the critical stream, *i.e.*, records which are critical for the publication figures.

Figure 1 depicts the distinct processes in micro-editing. After EDI or data entry from traditional paper forms, come sequentially

1. application of range checks,
2. automatic editing, for instance using the Fellegi-Holt rationale,
3. filtering the 'non-critical stream', for which automatic correction has been successful and
4. applying computer-assisted micro-editing to the remaining 'critical stream'.

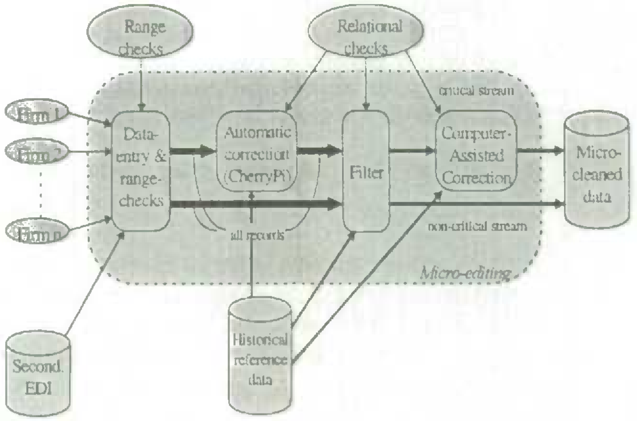


Figure 1. Data flow with micro-editing

Historical data can be used to detect erroneous fields in a record and for cold-deck imputation, both in automatic editing and in CAMAC. Section 2 gives details on automatic editing with the Statistics Netherlands program CHERRYPI , using parallel computation on several PC's in a network.

A score function or OK-index has been used to select the records which need CAMAC because of their influence on the publication figures or the size of suspected errors (Hidiroglou and Berthelot 1986; Van de Pol and Molenaar 1995). We presently consider the possibility to include the outcomes of automatic editing into this index (Figure 1). These outcomes should also be considered as a suggestion in the CAMAC process.

With macro-editing suspect micro-values are detected by looking for outliers in a set of homogeneous records. The set of homogeneous records is usually a publication cell, for instance an industry branch, size class combination. Generally the search will start at the macro-level, by comparing aggregates – provisional publication figures – with a prediction from history. At this level suspect aggregates can be selected. Subsequently, the set of records underlying this aggregate can be inspected to trace the field(s) that caused this suspicious change (Figure 2). If the aggregate is larger than expected, this meso-level analysis can be simply an inspection of high micro values or changes contributing to the total, that is a sort of records in a spreadsheet. A more general approach is to look for outliers in a multivariate distribution (Bethlehem and Van de Pol, 1997). Statistics Netherlands develops a graph-oriented program called MACROVIEW that offers a number of bivariate scattergrams for this purpose (section 3). This sort of analysis also produces information to be used in specifying edit checks (Figure 2).

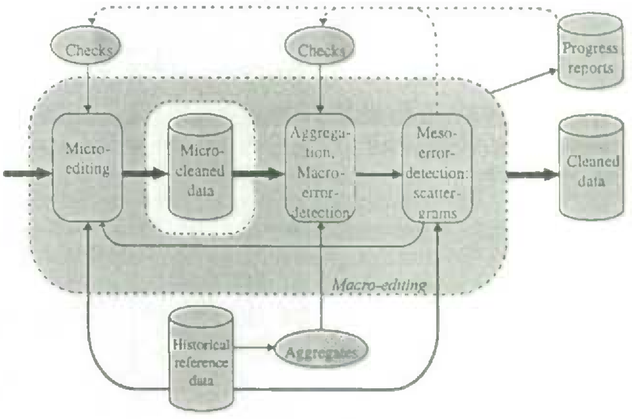

Checks
Checks
Progress reports
Micro-editing
Micro-cleaned data
Aggregation, Macro-error-detection
Meso-error-detection: scatter-grams
Cleaned data
Macro-editing
Historical reference data
Aggregates

Figure 2. Data flow including macro-editing

In order to control data editing there should be regular cumulative reports on the quantity of records processed and on the quality of corrections. Productivity can be monitored with the proportion of records processed in relation to the scheduled production time used. Engström (1996) mentions quality indicators such as the proportion of flagged records in time, the proportion of flags by branch, flags and corrections by variable and flags by edits. Important is also the effect of edits by variable in time. With these reports data editing can be monitored and checks can be adapted if necessary (Figure 2).

2. AUTOMATIC EDITING AND PARALLEL COMPUTATION WITH CHERRYPI

CHERRYPI is a program for editing numerical data automatically, given a set of checks. CHERRYPI has been written in Borland DELPHI 2.0, and runs under Windows '95. It is suited for parallel computing on several PC's in a network, but it can also run on a single PC.

The edits that can be handled by CHERRYPI can be written as

$$Ax \geq b, \quad (1)$$

where A is a constant matrix, $\boldsymbol{b}$ is a constant vector, and $\boldsymbol{x}$ is a vector corresponding to the values in a given record. The matrix A and the vector $\boldsymbol{b}$ together define the set of edits. For each stratum (publication cell) a distinct set of edits can be defined. To enter the set of edits for each stratum quickly and easily, a user-friendly interface has been developed and described in a user's manual (Bakker *et al.* 1997).

2.1 Error Localisation

The error localisation method of CHERRYPI is based on the Fellegi-Holt paradigm (*cf.* Fellegi and Holt 1976).

According to this paradigm a minimum number of fields should be imputed such that all edits are satisfied. Optionally, a weight may be assigned to each field indicating how trustworthy one considers the value of this field. The higher the weight of a field, the more trustworthy one considers the value of this field. In case weights are assigned to the fields CHERRYPI minimises the weighted number of fields such that all edits are satisfied.

To determine the fields that should be imputed according to the Fellegi-Holt paradigm, CHERRYPI applies the algorithm of Chernikova (*cf.* Chernikova 1964 and 1965; Rubin 1975). This algorithm generates a subset of the vertices of a set of linear inequalities. Chernikova's algorithm can be used to determine the fields that should be imputed (*cf.* Sande 1978). However, the basic algorithm is too slow to be useful in practice. Several improvements by Statistics Canada to speed up the error localisation (*cf.* Schiopu-Kratina and Kovar 1989) have been implemented in CHERRYPI . Due to these improvements CHERRYPI has become fast enough to be applied to middle-sized datasets. An application to records with 34 checks on 63 variables took on the average 2 seconds per record on a pentium 75, half an hour for every 1000 records.

In order to improve the quality of error localisation, the Chernikova algorithm in CHERRYPI is preceded by a step that makes extra use of equality checks and balance checks. These checks allow exact predictions of all variables involved. For those predictions that strongly resemble the observed figure, one of the following errors may be in order: sign reversal, double typing a key, skipping a key, interchanging a key or typing a key which looks similar or which is nearby on the keyboard. For any of these cases the field weight can be reduced appropriately so that the Chernikova algorithm will most likely choose that field for correction (Van de Pol, Bakker and de Waal 1997).

2.2 Imputation

CHERRYPI can apply deterministic and regression imputation; hot-deck imputation is under development. With deterministic imputation we mean that the imputation-variable, that is the variable which has been selected for imputation, can assume only one value such that all edits can be satisfied.

Regression imputation allows historical imputation (if the predictor is a historical variable), ratio imputation and mean imputation. After the imputation-variables have been imputed by means of regression imputation, the resulting record may still violate the edits. Therefore, in a second step the imputed values are modified slightly in order to satisfy all edits. In this way a consistent record is always obtained. For more information on regression imputation we refer to de Waal (1996).

In some branches of industry the dataset will be too small to provide a donor of the same size. We therefore consider to take all variables proportional to firm size before performing hot-deck imputation and to scale the imputed values with the firm size of the recipient record.

2.3 Parallel Computation on Several PC's in a Network

The Chernikova algorithm is not fast enough for processing datasets with hundreds of thousands of records on one PC. We therefore developed a version of CHERRYPI that can perform parallel computation on several PC's. This version is based on the Parallel Virtual Machine (P.M.) package, which originates from Oak Ridge National Laboratory. It is a C-program, which is presently free available on the Internet. (See www.epm.ornl.gov/p.m./pvm_home.html and Geist *et al.* 1994a, 1994b). Because parallel computation on a PC-network requires extra data transfer, it is especially useful when data transfer takes far less computer time than computation itself. This is the case with the Chernikova algorithm.

The parallel version of CHERRYPI runs as a master on one PC, which is in control of the data distribution. Slave PC's in the same network process most of the records. Master and slave PC's together form the 'virtual machine'. With the present Windows '95 implementation the slave PC's have to be started by hand and a 'demon' program has to be loaded locally, before the master PC CHERRYPI can activate and control CHERRYPI on the slave-PC's. Other operating systems, like Unix, enable also to switch on slave-PC's from the master PC. P.M. uses the network, but does not use the network software (NOVELL in our case).

A master-PC will not necessarily monopolise slave-PC's. In our implementation, which was designed by J. Kardaun, the users of the slave PC's remain on the contrary as free as they can be. Instead of priority 'normal', the local CHERRYPI jobs are given priority 'idle'. This means that the normal user of the slave PC will not notice any reduction in computation speed at his PC. Therefore parallel computation does not have to be performed in the night, but can just as well be performed during working hours. CHERRYPI jobs on slave-PC's will be visible on the Windows '95 task bar only and will take up just a little computer memory. Time-consuming screen savers should be switched off at the slave-PC's. More details are in Van der Horst (1997).

Table 1
First results with the parallel version of CHERRYPI

486-33 PC's	pentium-75 PC's	other use of slave PC's	time
1	–		47 min., 17 sec.
–	1		17 min., 23 sec.
1	2	none	10 min., 32 sec.
3	4	heavy	7 min., 44 sec.

The parallel version of CHERRYPI is still in an experimental phase. Table 1 shows results of tests on 1500 records. Considerable processing time reduction can clearly be obtained by parallel computation. The last line shows however that heavy use of slave PC's by local users will leave little time for the CHERRYPI slave jobs. Really heavy parallel jobs will have to continue in the night.

3. MACROVIEW

MACROVIEW is a program for graphical macro-editing. It has been written in Borland DELPHI 2.0, and runs under Windows '95. MACROVIEW gives a view on the data at three levels

1. the aggregate level of publication figures,
2. an intermediate (meso) level of bivariate distributions and
3. the micro level of records.

The aggregate level is to trace suspect publication figures, the intermediate level is to trace suspect data points in scattergrams and the micro level is to decide whether a field is actually in error and to correct it if necessary. The use of scattergrams at the intermediate level implies that MACROVIEW can only be used if at least part of the data is quantitative.

One of the key motivations which led to the development of MACROVIEW is that only corrections that matter for publication figures should be carried out. Scattergrams can be a powerful tool to find fields with important errors. Often previous measurements on the same units are used as predictors, but any variable with a strong relation to the suspect variable is good for this purpose.

Before we go into more detail on the three views of MACROVIEW some general remarks are in place.

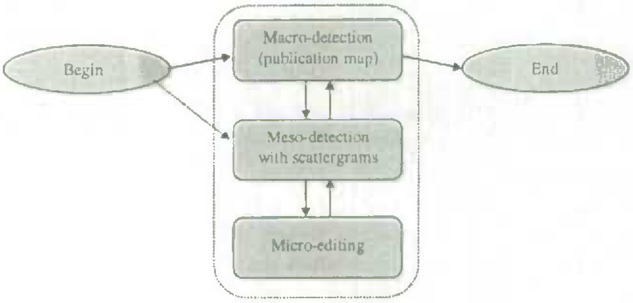

Begin
Macro-detection (publication map)
End
Meso-detection with scattergrams
Micro-editing

Figure 3. Graphical macro-editing needs smooth aggregation level changes

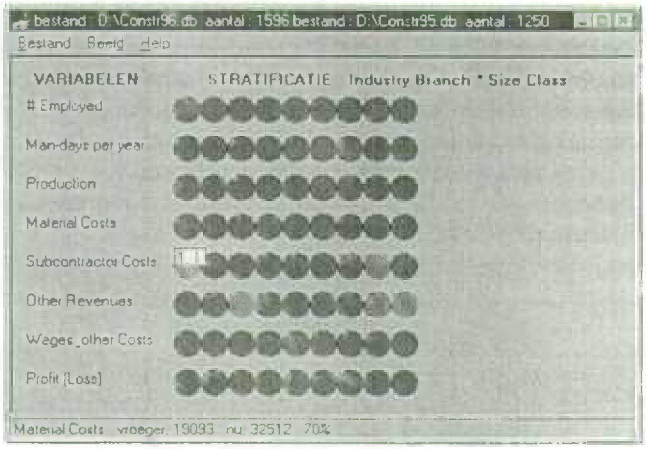
MACROVIEW is programmed in Delphi, a Pascal-based language which offers many build-in components for a graphical user interface. Despite the standard components, which are meant for *ad hoc* programming, we try to make a fairly general program for graphical macro editing.

An *ad hoc* program would only have one type of users, the data editors, specialists in some branch of industry. Their task is to correct errors and to localise them by zooming into and out of the micro level (Figure 1).

Because MACROVIEW is intended to become a general program, applicable to several surveys, we also have a second kind of users, namely the system designers (data editing chiefs) who specify the setting in which the data editors will work. One of the system design tasks is supported by MACROVIEW, namely the publication map layout. For CHERRYPI and BLAISE such a task is the definition of edit checks to be used at the micro level.

3.1 The Publication Map Level

The most aggregated view on the data is what we call the publication map and what Esposito *et al.* (1994) call the anomaly map. This map gives information about historical anomalies in publication figures and allows users to select suspicious subsets of the data. To determine how plausible new figures are, the dataset is compared to some historical prediction, for instance last year's data. In this map each publication figure is represented by a 'radio button'. By default, the map is a cross-table with variables as rows and publication cells (for instance industry branch by firm size class) as columns (Figure 4). To save space only few variables and cells are displayed here. A system designer can move the buttons and the text in any desired pattern.


bestand D:\Constr96.db aantal : 1596 bestand : D:\Constr95.db aantal : 1250

Material Costs vroeger. 19093 nu. 32512 70%

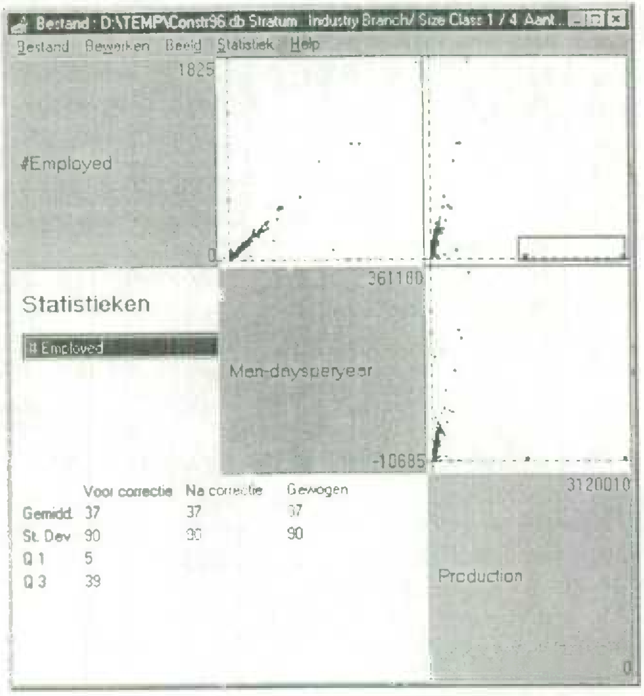
Figure 4. The publication map

The mouse which is not visible in Figure 4 points at material costs in publication cell 'Industry branch 1 and Size class 1'. The line at the bottom displays figures for the cell that is pointed at: past year's costs, *f* 19,093, present year's costs, *f* 32,512, and the percentual increase, +70%. This publication cell is coloured red, because the increase is larger than the threshold (here 50%). In case of a decrease the colour is purple, cells with a below-threshold change are green and empty cells are blue. In a similar program, called GEAQS, nine colour shades represent varying degrees of change (Weir 1996).

3.2 The Scattergram Level

Next, the intermediate level of scattergrams can be reached by double clicking on a publication figure. As Figure 3 shows, MACROVIEW can also be used without the publication map. This is useful when all the records should be plotted in the graphs instead of only those from a suspect publication cell.

The user should select scattergrams of the suspect variable with its best predictors. An outlier occurs when the predictors are reported consistently, whereas the suspect variable is far from the expected value. Localisation of outliers in these graphs is a powerful way to localise errors (Figure 5).



	Voor correctie	Na correctie	Gewogen
Gemidd.	37	37	37
St. Dev	90	90	90
Q 1	5		
Q 3	39		

Figure 5. Scattergrams to localize errors

Some of the outliers in Figure 5 have been 'brushed', that is marked by pulling the mouse diagonally over them, thus creating a box in one of the graphs. As a result, points representing the same records in the other graphs are also coloured red. This is to find out whether outliers in one variable are also outliers in other variables. When a record is regular with respect to all relations, except those concerning one specific variable (production in Figure 5), that variable may be in error.

Selection of suspect records can also be done statistically. The Mahalanobis distance is used to rank records. This distance d_i summarises the deviations of score x_{ij} on variable j from mean μ_j for all variables $1,\ldots j,\ldots J$ involved, standardising with the population covariance matrix $\sum$ of these J variables. With $\boldsymbol{x}_i$ the column-vector containing $x_{i1},\ldots x_{ij},\ldots x_{iJ}$ and μ the corresponding column-vector of means the Mahalanobis distance can be written as

$$d_i = \sqrt{(\boldsymbol{x}_i - \mu)' \sum\nolimits^{-1} (\boldsymbol{x}_i - \mu)}. \qquad (2)$$

This distance can be interpreted as the eccentricity of i, that is the length of the $\boldsymbol{x}_i - \mu$ vector in a J–dimensional space with standard deviation 1 on each dimension. It should be noted that the Mahalanobis distance, by standardising the selected variables, puts equal weight on each variable.

In order to get a robust estimate of μ and $\sum$ from data that may still hold erroneous outliers, an iterative procedure is used, which gives less weight to outliers (Buijs 1997). In the estimation formulas for a publication cell with n records

$$\hat{\mu}^q = \sum_{i=1}^{n} w_i^q \boldsymbol{x}_i \Big/ \sum_{i=1}^{n} w_i^q \text{ and}$$

$$\hat{\sum}{}^q = \sum_{i=1}^{n} (w_i^q \boldsymbol{x}_i - \mu)(w_i^q \boldsymbol{x} - \mu)' \Big/ \sum_{i=1}^{n} w_i^q \qquad (3)$$

for iteration q an outlying record i gets a Huber weight

$$w_i^q (d_i^{q-1}) = \begin{cases} 1 & \text{if} \quad d_i^{q-1} < c_J \\ c_J \big/ d_i^{q-1} & \text{if} \quad d_i^{q-1} \geq c_J \end{cases} \qquad (4)$$

which is inversely related to the Mahalanobis distance in the previous iteration. Iterations are started with $w_i^1 = 1$. As convergence criterion $\max_j |\hat{\mu}_j^q - \hat{\mu}_j^{q-1}| < 10^{-10}$ is used. The constant c_J is the 95% point of the χ_J^2 distribution with J degrees of freedom. This is a good choice when the x-variables are distributed multivariate normal, because in that case the Mahalanobis distance is χ_J^2 distributed (Little and Smith 1987).

When the statistical way to localise errors has been selected, and the user has specified the variables to use, all records with $d_i > c_J$ will be marked red in the scattergrams.

Data editors get information on the value range of each plotted variable and they may zoom into and out of the scattergrams. Moreover below the scattergrams some statistics are given to improve insight into the effects of data editing: the mean, the standard deviation and the values at the 25% and 75% points of the suspect distribution.

3.3 The Micro-edit Level

During micro-editing the user sees immediately the effect of corrections. After the correction has been affirmed (√) the second column down left in Figure 5, labelled 'na correctie', will show the new average. The third column is for statistics, weighted with some raising factor. Affirmed corrections are also displayed immediately in the scattergrams, marked with a deviant colour (blue). Furthermore the colour of the relevant publication figure in the anomaly map is updated when a correction has taken place.

The micro-editing environment appears as soon as the scattergrams have been requested. The first one of the brushed records is shown in a relatively simple data grid, a window with all variables of a record in one row. The values of a previous occasion may be added below the current values for reference. For data sets with few variables all variables are visible together, but for data sets with many variables scrolling is needed.

Corrections are stored in a log-file. This log-file enables restoration to a previous situation. More importantly, the log-file can be tabulated in various ways: number of corrections by variable, corrections by edit checks, the sum of the absolute value of corrections by variable, *etc*., (Engström 1996). This information is useful in trimming the data editing process.

4. FUTURE DEVELOPMENTS

Both CHERRYPI and MACROVIEW are still under development and are not used in production yet. CHERRYPI has been applied to Annual Construction Survey data with encouraging results. We presently prepare application of CHERRYPI in the Survey of Environmental Costs, which is scheduled for production in 1998. On the theoretical side, we will examine cutting plane algorithms, which may be more efficient than Chernikova's algorithm in combining categorical and numerical edit checks (de Waal 1997).

MACROVIEW is developed in co-operation with representatives from five distinct surveys. The zero-version has been tested in the usability lab of Statistics Netherlands. This version appeared useful for surveys with few variables and a limited number of errors. More complicated surveys need a better support for the user at the micro level. To remedy this, we intend to make BLAISE computer-assisted micro-editing available from MACROVIEW. This would give access to hard and soft error messages, among other things. Another envisaged extension is the possibility to mark scattergram points that have some characteristic, like for instance the administrative agency that delivered the records.

At present CHERRYPI reads ASCII data, whereas MACROVIEW is PARADOX oriented. Extension to other data formats like ACCESS, ORACLE and BLAISE will be considered when required for specific applications.

REFERENCES

Bakker, F.J., Been, G., van der Horst, J.P., van de Pol, F., and de Waal, T. (1997). Het gebruik van CHERRYPI. Statistics Netherlands, Voorburg.

Bethlehem, J., and van de Pol, F. (1997). The Future of Data Editing. Research paper. Statistics Netherlands, Voorburg. Eds., Cooper *et al.* To appear in: Computer assisted survey information collection (tentative title) New York: John Wiley.

Buijs, A. (1997). Graphical editing: an implementation. Statistics Netherlands, Voorburg.

de Waal, T. (1997). Cutting plane algorithms for optimal automatic error localization. Statistics Netherlands, Voorburg.

Engström, P. (1996). Monitoring the editing process. Conference of European Statisticians, Work Session on Statistical Data Editing, Voorburg, the Netherlands, Statistics Sweden.

Geist, A., *et al.* (1994a). PVM: Parallel Virtual Machine, A user's guide and tutorial for networked parallel computing. *The MIT press*, Cambridge.

Geist, A., *et al.* (1994b). PVM3 user's guide and reference manual. *The MIT press*, Cambridge.

Granquist, L. (1994). Macro-editing – A review of methods for rationalizing the editing of survey data. In United Nations, *Statistical Data Editing, Volume 1: Methods and Techniques*, Statistical Standards and Studies no.44, Geneva: United Nations Statistical Commission and Economic Commission for Europe, 111-126.

Granquist, L. (1995). Improving the traditional editing process. (Eds.) B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, and P.S. Kott. In *Business Survey Methods*, New York: John Wiley, 385-401.

Granquist, L., and Kovar, J.G. (1997). Editing of survey data: How much is enough? Eds., L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwartz and D. Trewin. In *Survey Measurement and Process Quality*, New York: John Wiley, 415-435.

Hidiroglou, M.A., and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology* 12, 73-83.

Little, R.J.A., and Smith, P.J. (1987). Editing and imputation for quantitative survey data. *JASA*, 82, 58-68.

Pierzchala, M. (1990). A review of the state of the art in automated data editing and imputation. *Journal of Official Statistics* 6, 355-377.

Van de Pol, F., Bakker, F., and de Waal, T. (1997). On Principles for Automatic Editing of Numerical Data with Equality Checks. Statistics Netherlands, Voorburg.

Van de Pol, F., and Molenaar, W. (1995). Selective and automatic editing with CADI-applications. Ed. V. Kuusela. In Essays on Blaise 1995; *Proceedings of the third International Blaise User's Conference*, Helsinki, Statistics Finland, 159-168.

Van der Horst, G. (1997). CherryPi Parallel; een parallel gaafmaakprogramma. Statistics Netherlands, Voorburg.

Weir, P., Emery, R., and Walker, J. (1996). The graphical editing analysis query system. Energy Information Administration, Washington, *1996 Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

MEASURING AND REPORTING THE QUALITY OF SURVEY DATA

D. Kasprzyk and G. Kalton[1]

ABSTRACT

Survey estimates are subject to sampling and nonsampling errors, the latter including nonresponse, noncoverage, measurement, and processing errors. This paper reviews methods for measuring these aspects of survey quality and discusses how the results should be communicated to users of survey data.

KEY WORDS: Sampling error; Nonsampling error; Survey quality; Survey reporting.

1. INTRODUCTION

Estimates derived from sample surveys are affected by sampling and nonsampling errors. Data collection organizations strive to build quality into their surveys by minimizing these errors subject to time and funding constraints and by promulgating standards in survey processes. For example, Statistics Canada (1987) has defined a set of "quality guidelines" for the production and maintenance of quality in data collections and the U.K. Government Statistical Service (1997) has developed guidelines in the form of a checklist of questions relating to individual areas of the survey process. Others suggest building "continuous quality improvement" systems into data collections (Morganstein and Marker 1997). These approaches advocate the use of statistical indicators of survey quality for measuring and reporting on the quality of survey data.

The premise underlying this paper is that information on the magnitude and extent of errors from different sources is important for the users of survey data (to inform their understanding of their analyses) and for methodologists (to assist in the development of improved methods for future surveys). Gonzalez *et al.* (1975) provide an excellent early review of the standards for the presentation of survey errors employed at the Bureau of the Census. In this paper we note that survey organizations may develop many different measures of survey data quality for each source of error, and that they communicate their knowledge of error in a variety of ways depending on the nature of the data being released.

The paper provides a general review of methods for measuring quality for various error sources – sampling, nonresponse, coverage, measurement, and processing errors – and discusses how and to what extent these indicators are communicated to the user community. From a broad perspective, quality of statistical data can be viewed as covering issues of accuracy, relevance, timeliness, accessibility, and cost-efficiency. This paper, however, is concerned only with accuracy and the communication of information on accuracy.

2. SAMPLING ERROR

Sampling error is probably the most well-known source of survey error. The reporting of sampling error for survey estimates has priority for all statistical agencies, and this is reflected in all statistical products, reports and data sets. For any survey based on a probability sample, data from the survey can be used to estimate the standard errors of survey estimates. Nowadays, the standard error of nearly any estimate can be readily computed using software that takes account of the survey's complex sample design.

The challenge that occurs with the computation of standard errors arises because of the multi-purpose nature of surveys. Surveys produce a very large number of estimates, often set out in many tabulations. The computation of standard errors for all the survey estimates, and also for differences between estimates, is a major undertaking; therefore, the direct calculation of standard errors for all estimates has not been a common practice. Rather calculations are performed for a number of key estimates and generalized variance functions (GVFs) are developed for predicting the standard errors of the remainder (Wolter 1985). These GVFs provide only approximate estimates of standard errors, and sometimes the approximations are not satisfactory (Bye and Gallicchio 1988). The advances in computing power in recent years make it more feasible to compute standard errors for all survey estimates, and it now appears that less reliance is being placed on GVFs.

Sampling error is often the only error source presented when reporting survey estimates. With press releases, for example, it is common to specify sampling error without mention of other error sources. When sampling errors are not reported by the media, they may well report sample size, presumably as an indicator of sampling error. Sampling

[1] Daniel Kasprzyk, U.S. National Center for Education Statistics, 555 New Jersey Avenue, N.W., room 422H, Washington, D.C. 20208-5574. Graham Kalton, Westat and Joint Program Survey Methodology, University of Maryland, U.S.A.

error may be communicated in a number of different forms, *e.g.*, standard errors, coefficients of variation, and confidence intervals. When reporting confidence intervals, the confidence level ought to be clearly stated. The Census Bureau recommends that confidence intervals and sampling variability be explained and that these explanations relate explicitly to the text that presents the estimates (U.S. Bureau of the Census 1995).

As has been noted above, generalized variance functions (GVFs) are often used for approximating standard errors. Survey reports frequently contain an appendix detailing the GVFs, and instructing readers in their use. Other reports contain an appendix that provides the standard errors for all the estimates in the tables in the main report. Both these procedures have the advantage of producing tables in the main report that are uncluttered with standard errors. There is, however, a concern that relegating the standard errors to an appendix leads them to being ignored by users.

Users' guides generally advise users analyzing public use files on methods for estimating standard errors. This advice may include both methods for computing the standard errors directly and methods for approximating them from GVFs. In general, it seems preferable to recommend direct computation. To facilitate the use of this approach, the users' guide should provide detailed instructions on its application and the data file should contain appropriate sampling error codes based on strata and primary sampling units (PSUs), suitably masked to protect confidentiality.

3. NONRESPONSE ERROR

Nonresponse is the most visible and well known source of nonsampling error. Nonresponse rates are frequently reported and are often viewed as the first area requiring study to assess the potential for bias in survey estimates. Failure to achieve a high response rate influences perceptions of the overall quality of the survey.

Unit nonresponse rates are a common output of the data collection process and are calculated differently for different purposes (Lessler and Kalsbeek 1992; Gonzalez *et al.* 1994; CASRO 1982). Both unweighted and weighted (weighted by the inverse of the probability of selection) response rates are easily, although not uniformly, calculated. These rates are only indicative of survey data quality, since they do not provide estimates of the level of bias associated with survey estimates. Studies collecting data from both respondents and nonrespondents are necessary to estimate bias. Moreover, an examination of bias should ideally take into account any nonresponse adjustments made to attempt to reduce the bias.

The complexities of the survey design often make calculation and communication of response rates confusing and potentially problematic. Nonresponse in surveys with several stages of data collection can easily be misrepresented. For example, surveys of school teachers often rely on two stages of data collection. First, sampled schools are asked for lists of their teachers; then a sample of teachers is selected from the lists provided by the schools. Response rates may be calculated at each stage and the overall rate is the product of the response rate at each stage (Scheuren *et al.* 1996). The overall response rate is the key measure, although it is important to report both stages. Response rates for random digit dial surveys are computed in a variety of ways, the major issue being how unanswered phone calls are treated (Massey *et al.* forthcoming). Response rates for panel surveys have several dimensions and the rate calculated may depend on the application – a rate for each wave, a cumulative rate for the panel, or a series of rates to describe the patterns of wave nonresponse (Jabine *et al.* 1990).

Other nonresponse indicators, such as the completion rate, refusal rate, and the out-of-scope rate, can be readily calculated to provide information about the quality of survey operations. An understanding of these components of nonresponse is important for both the data collector and the analyst. For example, a study of the components of nonresponse may help to identify ways in which the data collection procedure may be improved.

Other data quality indicators relating to nonresponse can be identified, but they are not usually reported in a formal way. Statistics providing the average number of contact attempts and the extent of refusal conversions help to indicate the level of effort expended to achieve response. The calculation of response rates by different levels of geography (region, state, metropolitan status) and by selected economic and demographic characteristics of the population (*e.g.*, total sales, size of firm, proxies for income, and race/ethnicity) are important for subnational and subpopulation analyses. In the same spirit, nonresponse adjustment factors, when conveniently displayed, can provide valuable information to an analyst.

Special studies are needed to estimate the bias due to nonresponse. Some studies measure and analyze variables common to both respondents and nonrespondents, relying on variables found on the sampling frame (Khare *et al.* 1994; Scheuren *et al.* 1996). Occasionally, a subsample of nonrespondents is followed up intensively to gather a few key variables; with this type of study, estimates of bias can be computed for some actual survey estimates. Such studies are important and may even be critical for surveys with high nonresponse rates, but their value is often limited because of a low response rate for the followup.

As with unit nonresponse, with item nonresponse the reporting of response rates is important and measuring differences between respondents and nonrespondents is helpful for understanding the limitations of some analyses. An indicator of quality related to item nonresponse is the provision of imputation flags on public use data sets. While this idea is gaining in acceptance and standardization, it is not implemented universally.

Unit and item response rates are commonly computed by data collection organizations and these rates tend to be treated as a proxy for survey data quality – more so than almost any other indicator one might propose. It is therefore somewhat surprising that response rates are not always reported in analytic publications. Understandably, these rates are generally not reported in short format analyses and press releases because of the constrained nature of the publication. Still, this is a debatable policy; a good case could be made for some general information on nonresponse rates, or at least a reference to a survey document containing such information, being provided in such publications.

The importance and complexities of nonresponse indicate the need for a comprehensive treatment, as can be provided in survey documentation, user guides, and quality profiles. These venues can include an in-depth discussion of the issue, illustrating differences between respondents and nonrespondents, and assessing the effect of nonresponse on survey estimates, thereby serving as valuable reference documents.

4. COVERAGE ERROR

Coverage error is the error associated with the failure to include some population units in the frame used for sample selection (undercoverage) and the error associated with not identifying units represented on the frame more than once (overcoverage). The principal source leading to coverage error is the sampling frame itself. Thus, it is important for the user to have some information about the quality of the sampling frame and its completeness for the intended target population.

Measurement methods and data quality indicators for coverage error rely on methods external to the survey operations. A frequent indicator of the magnitude of coverage error is the "coverage ratio". In this case, survey estimates are weighted to agree with independent current estimates of the survey population. With household surveys, at the final stage of weighting sample persons are usually weighted to age, race, and sex population totals and the reciprocals of these final weighting factors are known as coverage ratios. When properly displayed, the ratios provide an easy way to understand the relationship between the population subgroup estimate from the survey and a population total from some other nonsurvey source (administrative records, censuses, lists, analytic techniques).

Coverage indicators can also be obtained from record check studies in which current listings of the population of interest are taken from independent sources and compared record for record with the sampling frame being evaluated. There are a number of difficulties associated with this approach, such as matching problems and obtaining current independent sources. However, an examination of individuals or entities in selected categories not accounted for on the frame is useful for understanding analytic limitations and for suggesting frame improvement strategies. See, for example, Owens (1997).

Other methods to measure coverage exist. Reinterview studies may, for example, be used to obtain a more complete and accurate listing or identification of units which may then be compared with the larger data collection results from the same geographic areas. Indicators of quality are the process statistics resulting from the study – per cent of person/entities missed, number of duplicates, *etc*. Finally, another indicator of coverage comes from the reporting of a precise definition of the study population including population exclusions, such as the homeless, institutionalized, and the nontelephone households.

Coverage has received substantial attention in the United States over the past ten years because of the decennial census and the undercount of minorities. Despite this, however, coverage error is rarely mentioned, except as a possible source of error, in venues such as press releases, *Census Briefs*, *Issue Briefs* and *Statistics in Brief*. These products are short and, given the page limitations, there is little opportunity to document this error source. However, noncoverage rates can often be substantial for certain population subgroups, and may be as large as, or larger than, nonresponse rates. The limited reporting of coverage error has led to it being a somewhat neglected error source.

Analytic reports often provide some information on the nature of coverage error in sections or appendices called "technical notes" or "source and accuracy statements". Since these reports usually provide more sophisticated analyses and typically do not have page limitations, they ought to contain some minimum level of information. Occasionally an overall reporting of a coverage rate and a short statement naming the affected subgroups is given (Norton and Miller 1992). Key analytic variables most likely affected by coverage error should be discussed, but rarely are.

The cost and complexity of measuring coverage error suggests that special studies on this subject ought to be reported in special technical reports where detailed tables can provide estimates of undercounts on many characteristics. Unfortunately, such studies are often reported substantially later than the initial results and, therefore, are not viewed as useful to the policy makers who use the survey data.

5. MEASUREMENT ERROR

A major concern about the quality of survey data is the accuracy of the questionnaire responses. Measurement errors may arise in respondents' answers to survey questions for a variety of reasons, including misunderstanding of the meaning of the question, failure to recall the information correctly, failure to construct the response correctly (*e.g.*, summing the components of the amount requested), and social desirability bias. In interview surveys, measurement errors may arise from the interviewers who may cause respondents to provide inaccurate reports by asking the question or probing incorrectly, misinterpreting responses, or making errors in recording responses.

Measurement errors are the most difficult aspect of survey data quality to quantify. Special studies are required, and they are often expensive to conduct. A key distinction in categorizing measurement error studies is between those that attempt to assess measurement bias and those that are concerned only with measurement variance. Studies of measurement bias need to obtain measures of "true values" with which the survey responses can be compared. Studies of measurement variance investigate only the variability of responses across repeated applications of the survey process, sometimes to estimate the variable error associated with a particular source (*e.g.*, interviewer, designated respondent, or question form).

A common method for studying measurement bias is by means of a record check study. In such studies survey responses are compared with the values in some record system. Record check studies may be classified into three types: reverse, forward, or full design record checks (Groves 1989). A reverse record check study identifies over-reports by checking (validating) survey responses against administrative record data. A forward record check study identifies under-reports by first selecting cases from an administrative record source then conducting survey interviews to see if the record information is reported. The full design record check combines the two designs so that both under-reports and over-reports can be identified. Record check studies are valuable tools for studying measurement errors provided that records exist for the survey item under study and that access can be gained to them. However, the problems of matching errors and errors in the records should not be overlooked.

Another common design for a measurement error study is a reinterview survey in which a sample of survey respondents is reinterviewed (Forsman and Schreiner 1991). One form of reinterview survey simply reinterviews respondents using the same procedures used for the main survey. Under the assumption the two interviews are independent, this design can be used to estimate measurement variance. In a variant of this design, the second interviewer is asked to reconcile any differences in the answers given on the two occasions, with the aim of determining the "true values". In another variant, the second interview is different in nature, aiming to obtain "true values" by a very detailed interviewing procedure, perhaps conducted by a highly trained interviewer.

A third design uses different survey procedures for different replicates of the sample. Thus, different interviewers may be assigned to different replicates to measure interviewer variance. This method is widely used to measure the effects of variations in question form or different modes of data collection.

Finally, mention should be made of the various cognitive research methods that are widely used in questionnaire development. The use of such techniques as think-aloud, detailed probing and behavior coding can provide valuable insights in the quality of survey responses.

Measurement error studies are generally substantial projects reported in separate reports and in the survey methodology literature. Their findings are usually too complex to fully document in brief accounts, such as the sources of error appendices that appear in many survey reports. They are often not fully covered in the users' guides that accompany public use data files. As a result, there is a risk that the results of these studies may not be adequately communicated to users. In this situation, quality profiles provide a means for reporting information from measurement error studies for a survey in a single place, thus enabling a user to find out about them and identify the full reports if more detail is needed.

6. PROCESSING ERROR

After the survey data are collected, a set of processes takes place to handle the reported data and convert them to consistent machine-readable information. The processes include the receipt and control of questionnaires, manual editing and manual coding, data entry, and machine editing and imputation. They tend not to be well-reported or even well documented in survey materials; they also do not have a prominent role in the research literature (Lyberg and Kasprzyk 1997). They do, however, play an important role in transforming the raw survey data into an analytic format.

Processing errors include data entry, coding, and editing errors. Data entry errors may be measured through the use of a quality control sample, whereby a sample of questionnaires is selected for re-entry and an indicator of the quality of the operation, a keying error rate, is determined. Often other strategies are developed to achieve acceptable keying error rates. For example, in the American Housing Survey the work of new data keyers is verified (completely rekeyed as a check for errors) until the error rate is at or below a certain level, at which time only a sample of questionnaires is checked (Chakrabarty and Torres 1996). Keying error rates continue to be important with new technological advances as interviewers with little data entry experience now enter information in the computer assisted personal interviewing environment. Dielman and Couper (1995) report keying error rates of 0.095% in this new survey environment.

Coding is the process of classifying open-ended responses into predetermined categories. It is often a difficult task with many opportunities for individual coders to introduce their personal biases into the coding structure. Indicators of quality for this source of error can be obtained by selecting a sample of the coders' work to determine the number of errors – so that error rates can be calculated. Another indicator of quality is a comparison of all or some of a coder's coding decisions with another coder's decisions on the same cases and reporting a statistic which measures agreement on coding decisions.

Survey data editing is a procedure for identifying errors created through data collection or data entry using established edit rules. Data quality indicators that relate to errors identified in editing include tables of the number of edit changes for each item (Gruber *et al.* 1996) and tables of edit changes by reason for changes (U.S. Energy Information Administration 1996). Another indicator is a questionnaire rejection rate, whereby invalid entries in the questionnaire are identified, rejected, and corrected by the professional staff.

The details and extent of processing error are rarely discussed in agency reports. Analytic reports identify it as a source of nonsampling error and generally note that quality control procedures were implemented to reduce such errors. Performance statistics generated by the data entry, coding, and editing processes are developed and used primarily to ensure that the processing meets the agency's performance objectives and to provide feedback to the survey operations team, including individual data entry staff and coders.

The details of the processing aspects of the survey operations are best described in survey documentation reports or special technical reports. Detailed reports for survey practitioners, methodologists, and "serious" analysts have the benefit of drawing attention to operations most people take for granted. The National Center for Health Statistics, for example, has drafted a report documenting the editing practices for all its survey programs (U.S. National Center for Health Statistics 1994). Alternatively, the synthesis of much of this information into one document, such as a quality profile, provides significant value to the broad survey user community. The quality profiles for the Residential Energy Consumption Survey (U.S. Energy Information Administration 1996) and the American Housing Survey (Chakrabarty and Torres 1996) are useful examples combining survey documentation and performance statistics.

7. COMPARISONS TO INDEPENDENT SOURCES

All survey programs conduct reviews of their data prior to their release. These reviews typically look at survey process data quality indicators, some of which have been mentioned above, such as unit and item nonresponse rates, and edit rates. Outliers are usually checked for accuracy and reasonableness. Many of these reviews are focussed at the record level and so constitute a micro record review of data quality. However, another kind of review is equally important and, while it is oftentimes implemented routinely, it is often not reported on. This review is what might be called a macro review; that is, the review compares survey estimates in the aggregate with comparable data from other sources to establish the "reasonableness" of the survey estimates. Comparable data sources used in these types of reviews include administrative data collected and maintained for nonstatistical reasons as well as other sample surveys.

Aggregate comparisons of survey data with data external to the survey afford a relatively easy and inexpensive way to understand the survey data's "fitness for use". Differences between two or more data sets or data series indicate that further review and evaluation may be necessary. However, data sets are rarely fully comparable on the critical dimensions of the study: population coverage, concepts being measured, the time frame for which data are collected, the method of collecting data (mail versus phone, for example), the questionnaires, and the recall period for data items may be different. Thus, analysts comparing estimates across two or more data sets should be cautious. This technique is often used with establishment surveys (U.S. Federal Committee on Statistical Methodology 1988), but the comparative results are reported infrequently.

Aggregate comparisons, with adequate descriptions of their limitations, can provide useful information about the quality of survey data. There does not appear to be a standard medium for the presentation of this type of information. Shea (1995), for example, in a technical appendix to an analytic report, compares poverty status as measured in the Survey of Income and Program Participation (SIPP) with poverty as measured in the March Supplement to the Current Population Survey (CPS). Vaughan (1988) provides comparisons of SIPP income data with administrative program data and with CPS income data in a conference proceedings volume. Nolin *et al.* (1997) report comparisons of the 1996 National Household Education Survey with a number of other data sets in a working paper. Jabine *et al.* (1990) synthesize results from a number of sources in the SIPP Quality Profile.

8. CONCLUSION

Agencies release information from and about a survey through a variety of means. Descriptive analyses featuring tabular presentations are often reported. More complex substantive analyses of the data are available in analytical reports, and recently a number of short format publications, such as press releases, *Issue Briefs*, *Census Briefs*, and *Statistics in Brief*, geared to unsophisticated readers have become available. Survey results are also becoming widely available on the Internet. Methodological/technical reports are occasionally used to describe the results of special studies, and quality profiles have become the format to synthesize results on all aspects of survey operations. Public use micro data are released through CD-ROM, diskette, Internet, as well as 9-track tape files. Thus, a myriad of report formats and data release venues are available.

The reporting of information about error sources is not uniform in any of the formats listed in the preceding paragraph. Press releases generally have some information about sampling error, while analytic reports usually provide information on sampling error and nonresponse error. Information about all other errors is often not reported. Short format publications may acknowledge the variety of error sources, but say very little about the magnitude of the errors. The recently introduced mode of dissemination of posting survey results on the Internet has considerable potential, as it develops, for providing users with links to information about error sources. Methodological/technical reports provide substantial detail about special error studies, but they become available only several years after the survey has been conducted. Quality profiles are, perhaps, the most useful way to synthesize information about repeated surveys, but they are not produced very frequently and rely on the availability of results from methodological studies. A brief review of the many publication formats indicates that error sources are not widely reported. In many instances, increased attention to reporting the magnitude and extent of error in surveys would be desirable.

ACKNOWLEDGEMENTS

This manuscript benefited greatly from the discussions of the Office of Management and Budget's Subcommittee on Reviewing and Reporting of the Quality of Federal Data, a subcommittee of the Federal Committee on Statistical Methodology. Members of the subcommittee are: Dale Atkinson, Barry Beckman, Ram Chakrabarty, Judy Conn, Charles Darby, Brian Harris-Kojetin, Howard Hogan, Nathaniel Howard, Nancy Kirkendall, Jim Knaub, Marilyn McMillen, Renee Miller, Chris Moriarity, Dennis Schwanz, Karl Seiber, Carolyn Shettle, Antoinette Ware-Martin, and John Wolken.

REFERENCES

Bye, B., and Gallicchio, S. (1988). A note on sampling variance estimates for Social Security program participants from the Survey of Income and Program Participation. *Social Security Bulletin*, 51(10), 4-21.

Chakrabarty, R.P., and Torres, G. (1996). *American Housing Survey: A Quality Profile*. Current Housing Reports, H121/95-1, Washington, DC: U.S. Government Printing Office.

Council of American Survey Organizations (CASRO) (1982). *On the Definitions of Response Rates*. New York: Port Jefferson.

Dielman, L., and Couper, M. (1995). Data quality in a CAPI survey: Keying errors. *Journal of Official Statistics*, 11, 141-146.

Forsman, G., and Schreiner, I. (1991). The design and analysis of reinterview: An overview. In *Measurement Errors in Surveys*, (Eds. P.B. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman,) New York: John Wiley, 279-302.

Gonzalez, M.E., Kasprzyk, D., and Scheuren, F. (1994). Nonresponse in federal surveys: an exploratory study. *Amstat News*, 208, Alexandria, VA: American Statistical Association.

Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. (1975). Standards for discussion and presentation of errors in survey and census data. *Journal of the American Statistical Association*, 70, 351(II), 5-23.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley.

Gruber, K., Rohr, C., and Fondelier, S. (1996). *1993-94 Schools and Staffing Survey Data File User's Manual, Volume 1: Survey Documentation*. (NCES-96-142), Washington, DC: National Center for Education Statistics.

Jabine, T., King, K., and Petroni, R. (1990). *Quality Profile for the Survey of Income and Program Participation (SIPP)*. Washington, DC: U.S. Bureau of the Census.

Khare, M., Mohadjer, L.K., Ezzati-Rice, T.M., and Waksberg, J. (1994). An evaluation of nonresponse bias in NHANES III (1988-91). *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 949-954.

Lessler, J.T., and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York: John Wiley.

Lyberg, L., and Kasprzyk, D. (1997). Some aspects of post-survey processing. In *Survey Measurement and Process Quality*, (Eds. L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz, and D. Trewin) New York: John Wiley, 353-370.

Massey, J., Cox, B., and O'Connor, D. (forthcoming). *An Investigation of Response Rates in Random Digit Dialing (RDD) Surveys*. Washington, DC: National Center for Education Statistics.

Morganstein, D., and Marker, D. (1997). Continuous quality improvement in statistical agencies. In *Survey Measurement and Process Quality*, (Eds. L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz, and D. Trewin) New York: John Wiley, 475-500.

Nolin, M.J., Collins, M.A., Vaden-Kiernan, N., Davies, E., and Chandler, K. (1997). *Comparison of Estimates in the 1996 National Household Education Survey*. Working Paper 97-28, Washington, DC: National Center for Education Statistics.

Norton, A.J., and Miller, L.F. (1992). *Marriage, Divorce, and Remarriage in the 1990's*. Current Population Reports, 23-180, Washington, DC: U.S. Government Printing Office.

Owens, S. (1997). *Coverage Evaluation of the 1994-95 Common Core of Data: Public Elementary/Secondary Agency Universe Survey*, (NCES-97-505). Washington, DC: National Center for Education Statistics.

Scheuren, F., Monaco, D., Zhang, F., Ikosi, G., Chang, M., and Gruber, K. (1996). *An Exploratory Analysis of Response Rates in the 1990-91 Schools and Staffing Survey (SASS)*, (NCES-96-338). Washington, DC: National Center for Education Statistics.

Shea, M. (1995). *Dynamics of Economic Well-Being: Program Participation, 1990 to 1992*. Current Population Reports, 7-41, Washington, DC: U.S. Government Printing Office.

Statistics Canada (1987). *Quality Guidelines*. Ottawa, Canada.

U.K. Government Statistical Service (1997). *Statistical Quality Checklist*. London: Office for National Statistics.

U.S. Bureau of the Census (1995). Recommendations for Improving Publications and Press Releases. Memorandum from C.P. Pautler, Jr. to R.D. Tortora and T.L. Mesenbourg, February 7.

U.S. Energy Information Administration (1996). *Residential Energy Consumption Survey Quality Profile*. Washington, DC: U.S. Department of Energy.

U.S. Federal Committee on Statistical Methodology (1988). *Quality in Establishment Surveys*. Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs.

U.S. National Center for Health Statistics (1994). Data Editing at the National Center for Health Statistics. Internal draft report available from K. Harris, Chair, NCHS Data Editing Committee.

Vaughan, D.R. (1988). Reflections on the income estimates from the initial panel of the Survey of Income and Program Participation. In *Individuals and Families in Transition: Understanding Change Through Longitudinal Data*, Washington, DC: U.S. Bureau of the Census.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

SESSION I-10

Pushing the Frontiers of Design and Estimation: Data from Multiple Sources

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

THE AMERICAN COMMUNITY SURVEY DESIGN ISSUES AND INITIAL TEST RESULTS

C.H. Alexander[1]

ABSTRACT

The American Community Survey (ACS) will be a large continuing survey of the U.S. population using a "rolling sample" design of the type described in Kish (1990). It is a mail survey with follow-up by telephone and personal visit of a sample of nonrespondents. The Census bureau's ongoing Master Address File operation provides the frame. The ACS collects information about the same topics as the census "long form" content sample questionnaire. After a period of testing and comparison to the 2000 census long form, the ACS sample will increase to about three million mailouts each year starting 2003, leading to replacement of the content sample in the 2010 census. This paper describes the ACS survey design, objectives, and estimation procedures along with the considerations that led to them. The major methodological issues in conducting the survey and assessing the quality of the data are outlined. Initial results from the 1996 tests in four counties are presented, along with plans for future research and testing.

KEY WORDS: Rolling sample; Small area; Survey coverage.

1. INTRODUCTION

The American Community Survey (ACS) is a rolling sample survey being developed by the U.S. Bureau of the Census as an eventual replacement for the decennial census "long form" survey that provides the detailed economic, social, and housing characteristics of communities throughout the U.S. The ACS will cover the same topics as the census long form, but instead of contacting about 17,000,000 addresses at one time, the ACS will mail to about 3,000,000 addresses each year throughout the decade.

The ACS design has two main estimation objectives:

(a) provide descriptive profiles for communities of all sizes with mean squared error (MSE) generally similar to the census long form estimates, but updated throughout the decade;

(b) provide a time series of annual estimates for communities well below the state level, to measure changing local conditions.

After Census 2000, the ACS will replace the census long form sample as the source of detailed estimates of the characteristics of small areas. There will still be a decennial census to get a population and housing "count."

There has been a long-standing interest in updating the census descriptive profiles (Melnick 1991 or Sawyer 1993) in part because of their role in allocating Federal funds to local areas. The interest in tracking changes for sub-state areas has increased because of recent political developments in the U.S. sometimes referred to as "devolution" of decision-making to state and local governments.

To meet these two objectives, the ACS uses a rolling sample design suggested by Kish (1990, 1981) with what Kish calls "asymmetrical cumulation" of the survey data, *i.e.*, cumulating different numbers of years of data for different geographic levels or different uses. For the first objective, to produce estimates comparable in variance and bias to census long form estimates, 5 years of data would be cumulated to get sufficient sample sizes. For the second objective, to track changes for states and sub-state areas, single-year estimates would be used with the understanding that the larger sampling variance would mean that only very large changes could be detected for the smallest domains such as census tracts. The ACS questionnaire will initially cover the same topics as the 2000 census long form. After 2002, additional topics might be added. Participation in the ACS by sample households is required by law.

The Census Bureau views the ACS as part of a larger "continuous measurement" program, including benefits for the statistical programs of other Federal agencies. Besides the direct ACS data collection, this includes:

(a) use of ACS information to improve the Bureau's intercensal demographic estimates, which are in turn used in weighting the ACS;

(b) use of the ACS field staff to help update MAF/TIGER, keeping it complete enough for an intercensal survey frame;

(c) use of the ACS data and sampling frame to improve the estimation from various household surveys conducted by the Census Bureau, such as the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP) or the National Health Interview Survey (NHIS);

[1] Charles H. Alexander, Bureau of the Census, Room 3705-3, Suitland, MD 20233, U.S.A.
This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

(d) development of model-based small-area estimates for specific characteristics using data from existing household surveys, administrative records and eventually the ACS; the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program is an example, as is the Bureau of Labor Statistics Local Area Unemployment Statistics (LAUS) program (Brown 1997).

The ACS is being introduced in stages to allow review of the data, refinement of the operations, and a comparison with the Census 2000 long form before proceeding with a full introduction of the new survey. The stages are as follows:

1996-1998	Demonstration and Testing Period (four sites in 1996, eight in 1997, nine in 1998)
1999-2001	Thirty-seven "comparison sites" with 5% annual sample
2000-2002	National comparison sample with overall rate of 0.7% annually
2003-on	Full introduction (three million addresses per year, including all counties).

2. ACS DESIGN AND OPERATIONS

2.1 Sample Design

The ACS uses a "rolling" sample design with each address being interviewed at most once in a 5 year period. Each year's sample addresses will be spread evenly across the 12 months of the year and, starting in 2003, across the entire nation. The sample is in general not clustered, although there may be some exceptions in areas with unusually high travel costs and in group quarters.

The sampling frame will be the Census Bureau's Master Address File (MAF). The MAF is being created for the 2000 census, but is being created early for the ACS test sites. It starts with the 1990 census Address Control File, which is linked to the TIGER geographic database. This is updated using Postal Delivery Systems Files (DSF) in areas where the DSF addresses can be geocoded based on a "city-style" house number and street name. In other areas, the MAF must be created by physically listing each block. After Census 2000, the MAF will be kept up to date by regular matches to the DSF, at least every 6 months, by additional listing in rural areas of high growth as identified by administrative records systems. The sample will be selected as a systematic sample from the MAF, including only addresses not selected in the previous four years.

2.2 Data Collection

For each monthly sample panel, the ACS starts by mailing a questionnaire to each address about 10 days before the start of the "mailout month." There is a "pre-notice" letter, an initial mail questionnaire, and a reminder card, one week apart. After about 3 weeks, a replacement questionnaire is mailed if no response has been received.

At the beginning of the following month, nonresponding addresses are assigned to telephone nonresponse follow-up. For addresses where the telephone number can be obtained from commercial directories, a telephone interview is attempted. Mail returns continue to come in during this second month; in 1996 about 19% of the telephone nonresponse follow-up group were removed from follow-up because of a late mail return.

The third month, any addresses still not interviewed are eligible for personal-visit follow-up. This includes addresses where no telephone number could be located, as well as addresses where there was a number but no interview could be obtained. One-third of these addresses are selected for follow-up by personal visit. Note that this includes most of the vacant addresses.

As an example, the March mailout panel has telephone follow-up in April, and personal-visit follow-up in May. In general, a new area introduced in a particular year starts with mailouts in November and December of the previous year, so that the normal pattern of follow-up work in January and February is in place by the time the year starts.

Roughly speaking, based on the results of the 1996 and 1997 tests, we expect a National average of about 70% of the sampled addresses to be completed by mail or telephone and about 1/3 of the remaining 30% to be selected for personal visit followup.

The mail returns undergo a clerical edit which includes determining

(1) if the form is missing enough responses to require a callback;

(2) if an initial write-in entry giving the number of persons at the address was inconsistent with the number actually included in the questionnaire;

(3) if more than 5 people were listed as living at the address, since the form only collects characteristics for 5 people.

The first condition is referred to as the "content edit" and the last two as the "coverage edit." If the form fails either edit, then there is a telephone callback that attempts to fill in all missing data and straighten out any coverage problems. The exception is that if the content edit finds that the form is completely blank, the case continues to nonresponse follow-up as if no form had been returned. In 1996, about half of the mail returns failed one or more edits; of these 96% gave a telephone number, and 92% of these had some further resolution from telephone followup.

2.3 Residence Rule and Reference Period

The residents of any sample address, and their characteristics, will be determined as of the time of data collection. This refers either to when the mail form is filled out or when the nonresponse follow-up interview takes place. The ACS currently uses a "2-month" rule to determine who is a "current resident" of an address. Anyone staying at the address more than two months is included as a current resident. People staying two months or less would also be included, unless they usually live somewhere else.

3. ACS ESTIMATION METHODS

3.1 ACS Weighting

The 1996 ACS estimates were weighted using fairly conventional survey methods combining features of the census long-form weighting and household surveys such as the Current Population Survey (CPS). The details are given in Alexander, Dahl, and Weidman (1997).

3.2 Edit and Imputation

The editing and imputation for the 1996 ACS was similar to that for the 1990 census long form, although there are minor improvements for specific data items, mainly associated with minor questionnaire changes that are also being considered for the 2000 census.

3.3 Variance Estimation

The sampling variances for ACS estimates have been estimated using replication methods similar to those used for the CPS, with reweighting of each replicate to account for the effect of population and housing controls. For some items, the variance was also estimated using the random group method used for the 1990 census. The results of the two methods are still being compared, but an initial review showed the results to be generally similar. As with the 1990 census long form, the ACS variance estimates do not include variance due to imputation of missing data. This will be remedied as soon as possible. The ACS variances will mainly be reported to the users in the form of "generalized variance functions" (GVF), which approximate the variance as a "design factor" times the corresponding variance from a simple random sample. Different groups of characteristics have different design factors. Work on the 1996 GVFs is still underway.

4. RELATIONSHIP TO OTHER FEDERAL GOVERNMENT SURVEYS

The ACS will provide a valuable "statistical infrastructure" to improve the estimates and operations for other Federal government household surveys. The ACS data can be used in weighting and sampling for these surveys, as census long form data have traditionally been used, but the ACS will be more up-to-date. The sample for other surveys can be supplemented with housing units having specific demographic or economic characteristics taken from recently interviewed ACS sample units. When the survey samples are redesigned after the 2000 census, the MAF will provide more flexibility for drawing additional sample between censuses. The ACS will also provide auxiliary variables to improve small area models such as those used for the Local Area Unemployment Statistics (LAUS) or Small-Area Income and Poverty Estimates (SAIPE) programs.

However, the ACS will not replace the need for the CPS to measure unemployment, the SIPP to measure the dynamics of income and poverty, or other special-purpose surveys such as the NHIS or the National Crime Victimization Survey (NCVS). The subject matter of these surveys requires specialized questions that are too complicated for a self-response mail questionnaire. The ACS design is not suitable for measuring month-to-month changes as does the CPS, nor for following people over a period of time as does the SIPP.

There has been concern about possible confusion between the more accurate national and state labor force estimates from the CPS, which is designed especially to measure labor force status and the corresponding ACS estimates. The Census Bureau and Bureau of Labor Statistics are working together to avoid such confusion (Brown 1997). The ACS will not produce monthly unemployment estimates. For annual estimates, the ACS will provide a set of "adjusted" unemployment and civilian labor force (CLF) responses that will give unemployment and CLF estimates agreeing exactly with CPS at the national level, and conforming more closely to the CPS estimates at the state level.

The potential benefits of the ACS for other Federal government surveys, and its limitations, are discussed in more detail in Brown (1997).

5. WHY THIS DESIGN?

5.1 The Alternative of Expanding the CPS

An alternative to having a separate mail survey like the ACS to produce intercensal small-area data would be to expand the CPS sample and have one large personal-visit survey to produce both annual small-area data and estimates of short-term change in labor force characteristics. A rolling sample design with approximately 2 million addresses per year (without sampling for nonresponse followup) could in theory serve both objectives. This is about four times the current monthly CPS sample, but would not dramatically reduce the variance of monthly change estimates because the high correlation between monthly estimates with the current CPS rotating panel design would be lost.

The unit cost of this design would be substantially higher than the current CPS. The CPS now uses a cluster sample, conducts most interviews by telephone using the phone numbers obtained on the first of eight interviews at each address, and has a shorter interview than a combined ACS/CPS survey would require. We think that a survey of this design would cost several times as much as the combined cost of the current CPS plus the projected $75 million annual cost of the ACS operations.

To make a compromise CPS/ACS design affordable, it would be necessary to give up on some of the objectives, sacrificing either top-quality monthly measurement of unemployment, or "long form" data for small areas such as census tracts. A review of uses of census data (Edmonston and Schultze 1995) showed that the long-form small-area estimates were necessary to meet legislative requirements, and the importance of the monthly unemployment estimates for economic decision-making is well established.

5.2 The Alternative of a Mid-Decade Census

With the failure of a mid-decade census to be funded for 1985 or 1995, this alternative was not extensively considered in the ACS development. For the purpose of updating census profiles for small areas, a mid-decade sample census is arguably as effective as the five-year averages proposed for the ACS. However, a quinquennial "snapshot" is not effective for monitoring year-to-year changes, the second major use of the ACS. This new use seems to have made the difference in obtaining support for the ACS.

5.3 Alternatives Relying Mainly on "Indirect" Estimation

Another alternative would have been to rely less on "direct" estimation from a large survey and more on "indirect" model-based methods combining information from administrative records and smaller surveys. The "smaller surveys" could be a modification of CPS and existing surveys, or they might include a smaller mail survey such as a much reduced ACS.

As mentioned in Section I, research on such methods is part of the Continuous Measurement program. As these methods develop and become accepted by data users, we hope some of the ACS sample will gradually be replaced by information from statistical models. However, the development of these methods is not far enough along to eliminate the need for large samples to produce estimates of a variety of characteristics for very small areas.

5.4 Why This Data Collection Design?

The uniform spread of the sample was needed to provide comparable estimates for all levels of geography each year. A precursor of the ACS with different areas in different years was previously explored (Herriot, Bateman and McCarthy 1989), but was rejected because of the difficulty in making comparisons across areas.

The choice of a mail survey with followup was based on experience with this design in the census. The particular multiple-mailings approach was selected based on research following the 1990 census. The limitations of the questions that can be asked by mail are not a barrier since the ACS objectives involve the topics covered by the census long form survey, which is also done by mail.

We decided not to have any clustering of the sample, rather than to use small "ultimate clusters" of, say, four adjacent addresses as does CPS. The unclustered design is more efficient for the mail survey. The relatively high mail response rate means that a large initial cluster would be needed to have an expected four, or even two, in a cluster for followup. We are still considering clustering of followup cases in remote areas by reassigning the followup cases in a particular area to the same month of interview.

The data collection procedures for the telephone and personal-visit followup interviews were adapted from those used for CPS and other Census Bureau household surveys. The subsampling rates were chosen considering the relative costs per interview for the mail, telephone, and personal-visit modes using the rule that the allocation should be inversely proportional to the square root of cost but rounding to whole-number sub-sampling rates (no subsampling for telephone and 1 in 3 for personal visit).

6. VARIANCE AND BIAS OF THE ACS ESTIMATES

The ACS was designed with certain tradeoffs in mind between sampling error, frequency of updating the data, various sources of measurement error, and issues of interpreting and using the data.

The intended tradeoffs for the smallest communities are as follows:

1) Sampling error: standard errors 1.25 times as large as the long form design for "typical" estimates for small areas;
2) Frequency of updating: annual rather than decennial;
3) Issues of interpreting and using the data: 5-year average rather than point-in-time;
4) Other nonsampling errors: roughly equivalent, with each design having relative strengths and weaknesses.

For larger domains, where 5 years of ACS sample is more than enough for many purposes, annual estimates or shorter averages can be used. In this case, sampling error and the interpretation issues concerning the multi-year averages are less important, and other nonsampling errors are relatively more important.

In presentations of the ACS plans, these intended quality tradeoffs have been described to potential ACS data users. Now that the preliminary 1996 data are available, we can begin to verify the first statement about the standard errors and the fourth about nonsampling errors. The remainder of this section gives a first look at the preliminary results. All these conclusions must be regarded as very tentative, since they are based on preliminary data, and a small non-probability sample of test areas.

6.1 Sampling Error: ACS Compared to 1990 Long Form

A preliminary comparison of the ACS standard errors for tracts, to the corresponding "1990 census standard errors" computed using the 1990 generalized variance function (GVF) assuming a 1-in-6 sampling rate, is generally consistent with the anticipated 1.25 ratio. However, the results show that a simplistic rule like "1.25 times as large" is only a general guide. The ratio is higher for items that are concentrated in the "nonresponse universe" – that portion of the population that would not respond by mail or telephone because of the ACS's use of sampling for nonresponse followup. The extreme case is for vacant units, which are almost all collected by personal visit. We are still working on summarizing this variance comparison and extrapolating to what can be expected from the 2003 ACS.

6.2 Evidence From the 1996 Test About Nonsampling Errors

Rather than attempt to quantify the net bias in ACS estimates in a "total error" analysis, we adopt the less ambitious approach of looking for evidence of specific quality problems and quantifying them separately, in some cases with only indirect measures.

The final weighted unit response rates were quite high, running at 98.2% in the 1996 test areas and 98.7% in the 1997 areas except Houston, from January through July 1997, and 97.2% in Houston, TX, which has many hard-to-enumerate areas. By contrast, long-form data were collected for only 91.5% of sample units in the 1990 census, although all the rest had the basic census count information collected. The ACS final response rates in Rockland, NY were uniformly high for all tracts although mail return rates varied dramatically. (See Salvo and Lobo 1997.)

Salvo and Lobo (1997) argue that a more valid comparison is to look at what proportion of households complete a specific questionnaire item, *i.e.*, they look at the combined effect of "unit" and "item" nonresponse. In the Rockland, NY test site, they found that item response rates for cases assigned to follow-up were uniformly at least as high for the ACS as for the 1990 long form and were substantially higher for some items. This was expected, as a product of having a permanent field staff.

We are not prepared to draw a conclusion about whether the ACS suffers from the same kinds of overall undercoverage of persons relative to census-based intercensal demographic estimates that is seen for the CPS and many other household surveys. Comparisons of ACS weighted population estimates prior to post-stratification come within a few percent of the intercensal population estimates for the test counties. The population estimates include an adjustment for undercount in the 1990 census (Table 1, row 1). However, it is possible that these results are overly favorable if the vacancy rate is under-estimated, as discussed below.

Table 1
Ratio of "Before PPSF" Estimate
Divided by "After PPSF" Estimate

	Site			
	Rockland County	Multnomah County	Brevard County	Fulton County
Total Persons	0.975	0.987	0.958	0.941
Race = "Black"	1.044	0.781	0.861	*
Race = "Other"	0.974	0.985	0.898	*
Hispanic Origin	0.930	1.162	0.849	*

* Sample too small for a reliable estimate

There is still reason for concern about differential undercoverage of non-white persons and persons of Hispanic origin. Results were mixed on this, with some sites showing undercoverage of these groups and others showing good coverage and in one case overcoverage (Table 1). The latter anomaly suggests that the race/Hispanic-origin breakdown in the population controls may not have fully captured changes since the census. (Alexander, Dahl, and Weidman 1997). If so, this demonstrated a situation where the ACS can provide information useful in improving the population controls. This analysis is complicated by differences in the race/origin categories between the ACS and the 1990 census, another reason that we are not prepared to draw conclusions without further study.

Comparisons of the 1996 ACS estimates with 1990 census estimates at the county level showed few differences that suggested methodological problems. The most salient concern is that the ACS had noticeably lower vacancy rates. Somewhat lower rates are expected because the residence rules reduce the number of "vacant-usual residence-elsewhere" situations, but there is a possibility that the vacancy rate could be underestimated because of the long period allowed for ACS follow-up, with units that are vacant at the time of mailout becoming occupied by the time of follow-up. This is being studied further.

Income, poverty, and other economic data were not ready in time for this analysis. There is concern that asking the income questions "for the last 12 months" throughout the year may give less accurate recall than asking for "the last calendar year" in April. A small test comparing "last 12 months" and "last calendar year" is being conducted at the end of 1997, and may give some insights about this issue.

Examination of changes for individual tracts has just begun. One dramatic error was found in a small tract in the Rockland test site where a geocoding error on the MAF caused the number of addresses to drop dramatically between 1990 and 1996. This illustrates the need to feed information about address problems from the ACS data collection process back to the MAF updating process. The ability to detect such MAF errors is a potential benefit of the ACS, but we do not yet have this system in place.

A fundamental question is how funds can be allocated equitably based on the most recent data when the best estimate for large cities may be for the previous year while for small places it may be for the previous 5 years. Two solutions have been suggested. The first is to use the longest average (5 years in most cases) for all areas. The second is to allocate funds to large areas using one year data and then within the large areas based on multi-year averages. The large areas could include collections of small rural counties in addition to large cities or metropolitan areas. This question also needs to be widely discussed.

6.3 Annual Average Data

The differences between the ACS and the long form survey associated with collecting data all year with a moving reference date, rather than in the few months after the census using a fixed reference date, may actually be more important than the use of multi-year averages. However, these differences have generated less concern among users, perhaps because annual averages are more familiar from other household surveys such as CPS. The 1999-2001 comparison sites have been selected to represent areas in which various sources of differences are expected to be especially important and we expect our understanding of these differences to grow as time goes on.

REFERENCES

Alexander, C.H., Dahl, S., and Weidman, L. (1997). *Making Estimates from the American Community Survey.* Presented at the 1997 Annual Meetings of the American Statistical Association.

Brown, S. (1997). *Potential Uses of Continuous Measurement in BLS Labor Force Programs.* Presented at the 1997 Annual Meetings of the American Statistical Association.

Center for Study of Social Policy (1995). *Making Decisions Count: How the Census Bureau's New Survey Could Transform Government.* 1250 Eye Street, NW, Washington, DC.

Chand, N., and Alexander, C.H. (1997). *Achieving Agreement in the American Community Survey and the Current Population Survey.* Presented at the 1997 Annual Meetings of the American Statistical Association.

Edmonston, B., and Schultze, C. (1995). Eds. *Modernizing the U.S. Census.* National Academy Press, Washington, DC.

Herriot, R.A., Bateman, D.B., and McCarthy, W.F. (1989). The decade census program – A new approach for meeting the nation's needs for sub-national data. *Proceedings of the American Statistical Social Statistics Section*, 351-355.

Kish, L. (1981). Population counts from cumulated samples. *In Using Cumulated Rolling Samples to Integrate Census and Survey Operations of the Census Bureau.* U.S. Government Printing Office, Washington, DC., June 26, 1981, 5-50.

Kish, L. (1990). Rolling samples and censuses. *Survey Methodology.* 16, 1, 63-79.

Melnick, D. (1991). The census of 2000 A.D. and beyond. *Review Major Alternatives for the Census in the Year 2000.* U.S. Government Printing Office, Washington, DC, August 1, 1991, 60-74.

Ott, K., Parmer, R.J., Loudermilk, C., McMillan, Y., Reilly, B., and Coughlin, T. (1997). Evaluation of the Census Bureau's Master Address File Using National Health Interview Address Listings. Presented at the 1997 Annual Meeting of the American Statistical Association, Survey Methods Research Section.

Salvo, J., and Lobo, P. (1997). The American Community Survey: Nonresponse Follow-up in the Rockland County Test Site. Presented at the 1997 Annual Meetings of the American Statistical Association.

Sawyer, T.C. (1993). Rethinking the Census: Reconciling the Demands for Accuracy and Precision in the 21st Century. Presented at the Research Conference on Undercounted Ethnic Populations, Bureau of the Census, May 7, 1993.

U.S. Department of Transportation, Bureau of Transportation Statistics (1996). *Implications of Continuous Measurement for the Uses of Census Data in Transportation Planning.* Washington, DC, April 1996.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

NONRESPONSE AND COVERAGE ADJUSTMENT FOR A DUAL-FRAME SURVEY

J. Waksberg, J.M. Brick[1], G. Shapiro, I. Flores-Cervantes, B. Bell and D. Ferraro

ABSTRACT

The National Survey of America's Families is a large household survey that combines an RDD telephone sample and an area sample to cover households with and without telephones. This paper focuses on the estimation strategy. The first aspect discussed is the adjustment for noncoverage. Since both frames are incomplete, weighting adjustments are made to compensate for the lack of coverage in each sample. The second aspect discussed is the method of handling nonresponse. Poststratification has a dual goal of reducing both the bias from both nonresponse and undercoverage and the variances of the estimates.

KEY WORDS: Dual-frame; Poststratification; Weighting adjustments; Bias adjustments.

1. INTRODUCTION

A number of critical changes in the operation of the social programs in the United States were made in 1997, primarily in devolving responsibility for implementing the programs from the federal to state levels and in revising the rules for entrance and retention in Aid to Families with Dependent Children, one of the key social programs. The National Survey of America's Families (NSAF) is part of *Assessing the New Federalism*, a multi-year Urban Institute study to assess the effects of these changes by tracking ongoing social policy reforms and relating policy changes to the status and well-being of children and adults. The major objective of the study is to assess the effects of the devolution of responsibility for major social programs from the federal government to the states focusing on health care, income security, job training and social services. The NSAF collected information on the economic, health, and social dimensions of well-being of children, non-aged adults, and their families in 13 states (and one sub-state area), and in the balance of the nation; the data will be intensively studied as part of the project. The 13 states, which account for about 50 percent of the country's population, were selected to provide variation in terms of size and geographic location, the dominant political party, and key baseline indicators of well-being and fiscal capacity. A sample of the balance of the nation is included so that national estimates can also be produced. Low-income families were oversampled because the policy changes of interest are anticipated to affect them most. The initial round of the NSAF took place in 1997 and a follow-up round is planned for 1999 or 2000. There are two rounds of case studies in parallel with the survey to provide a detailed understanding of the policy changes occurring in each of the 13 states. This study is being directed by The Urban Institute and Child Trends and is being funded by a consortium of foundations, led by the Annie E. Casey Foundation. Westat is responsible for sampling, data collection, processing and related activities.

This paper discusses the estimation strategy that is designed to reduce the bias due to undercoverage and nonresponse. Particular emphasis is given to an adjustment for undercoverage of households without telephones. This is important for the survey because block groups with very low rates of nontelephone households were excluded from the sampling frame.

1.1 Goals of the Survey

The survey was designed to provide social and economic characteristics of the total civilian, noninstitutional population of the U.S. under the age of 65, but there is a major focus on obtaining reliable estimates for persons and families below 200 percent of the poverty threshold, with emphasis on families with children under 18 years of age. For each statistic, estimates are needed for 13 selected sites and for the nation as a whole.

About one-third of U.S. households contain children under 18, and between 35 and 40 percent of these have income below 200 percent of poverty. So nearly eight households need to be contacted to locate each family with children below 200 percent of poverty. Since separate statistics were required for each of the 13 states and the balance of the U.S., a very large and costly screening effort was required.

One obvious way of reducing costs is to use RDD telephone sampling for screening. Unfortunately, RDD alone is likely to be subject to serious biases, since about 20 percent of families in poverty do not have telephones. The nontelephone households probably have even lower incomes than other poverty families, and their economic characteristics are also likely to be different. The first and major problem in the development of the survey design was how to keep costs within reasonable limits without introducing serious biases or large sampling errors.

[1] J. Michael Brick, Westat, 1650 Research Boulevard, Rockville, MD 20850, U.S.A.

Another issue in planning the survey was the integration of the 1997 and 1999 or 2000 surveys. We will not discuss this aspect of the sample design in this paper but concentrate on the sample and estimation of the 1997 survey. More details on the sample design of the NSAF, including some thoughts on plans for the followup survey, are contained in Waksberg *et al.* (forthcoming).

1.2 Main Features of the Sample Design

The data collection was mainly carried out in the first 10 months of 1997. Since the analytical plans call for separate estimates for each state, the sample was designed to include approximately equal sample sizes per state. The equal sample sizes refer to the number of low-income families with children rather than the number of households sampled.

The sample used two-frames: RDD to cover the approximately 95 percent of U.S. households with telephones and an area sample to represent households without telephones. The RDD portion of the survey was carried out through a list-assisted method for sample selection. The sample households were first screened for the presence of children under 18 years of age, and those containing only persons 65 and over were excluded. Households without children were subsampled for the adult sample. The households were further asked a brief question about their 1996 income. All households with reported income under 200 percent of poverty and a subsample of higher income families were retained for the longer, detailed interview. The subsampling rates for higher income families for households without children varied among the states.

The nontelephone households were selected via a stratified, multi-stage, area sample. The PSUs were ones that are commonly used in area samples – MSAs and counties or combinations of several counties. The number of sampled PSUs varied from 4 to 12 among the states, and with 18 PSUs for the "balance of the U.S." Area segments consisting of Census blocks or combinations of blocks constituted the second stage of sampling. Compact clusters were used, that is all households in the segments were in the sample but the actual, effective cluster size was much smaller since the households were screened for presence of telephone and only nontelephone households were interviewed.

To reduce the potentially high cost of screening for nontelephone households in parts of the nation with very few nontelephone households, the area sample frame was truncated to exclude block groups with very low proportions of households without telephones. 1990 Census data were used to determine cut-off levels for excluding block groups from the sampling frame. Cut-offs were determined on a state by state basis and chosen so that less than 10 percent of nontelephone households in the 1990 Census were excluded. The truncation of the sampling frame reduced the screening workload by about 55 percent. The cut-offs for truncation in each state, and their effects on the workload are shown in Table 1.

The decisions that determined the sample sizes for the survey are described in Waksberg *et al.* (forthcoming). An effective sample size of 800 poor children, 450 to 1,000 nonpoor children, 500 poor adults, and 300 to 700 nonpoor adults were specified for each state. Somewhat larger sample sizes were set for the balance of the U.S.

Table 1
Cut-off for exclusion of high telephone rate areas and effect on workloads

State/Area	Percent of cut-off for low percent nontelephone	Percent of nontelephone households excluded	Percent of all households excluded
Alabama	< 5	7.3	40.9
California	< 2	7.3	59.0
Colorado	< 3	8.8	57.9
Florida	< 3	9.1	54.6
Massachusetts	< 2	9.0	70.4
Michigan	< 3	9.8	59.9
Minnesota	< 2	9.1	60.7
Mississippi	< 8	9.7	35.9
New Jersey	< 2	5.6	66.8
New York	< 3	7.5	58.7
Texas	< 5	7.7	45.3
Washington	< 2	6.1	53.7
Wisconsin	< 2	9.2	59.5
Balance, U.S.	< 3	8.0	57.1

Source: 1990 Census Tabulations

The subsampling of nonpoor persons did create some problems in achieving the desired sample sizes and will have an effect on the precision of the sample estimates. This is a result of an important feature of the sample design in the RDD component. In order to encourage as high response rate as possible, the screening instrument was kept simple, with only a brief question on whether the income was above or below a particular level. More intensive probing for income was planned for the detailed interviews. It was recognized that the simple screening would not always provide the correct classification of families as poor or nonpoor, and a major factor in determining the actual sample sizes necessary to achieve the desired sampling errors was the expected extent of error in the screening. There is evidence from another survey that a simple question on income produces nontrivial response errors. The sizes of the screening errors in that survey and preliminary results from the NSAF are included in Waksberg *et al.* (forthcoming).

1.3 Current Status of the Project

Data collection activities have now been completed and the sampling weights and variance estimation procedures are proceeding. However, three aspects of the sample are disappointing. First, initial telephone cooperation rates were lower than anticipated. A number of steps were taken to improve the response rates but the final overall response rates in most states are lower than we expected, especially given a group of efforts to increase these rates.

A second unexpected outcome was a lower rate of nontelephone households in the truncated frame than was reported in the 1990 Census. A number of hypotheses have been developed and are being explored to explain this. We discuss some recently developed data that provide information on the reasons for the shortage of nontelephone households.

Third, the rate of misclassification of low-income households in screening was higher than anticipated in some states. As noted above, the higher the rate of misclassification the greater the variances of the estimates and the lower the yield in terms of poor children in the sample. This topic is discussed further in Waksberg *et al.* (forthcoming).

2. COVERAGE ADJUSTMENTS

In nearly every sample survey, coverage of the population is an issue that must be examined. Incomplete coverage typically results in biases in the estimates and the biases may be substantial if a large proportion of the population is not covered and the covered and uncovered populations have different characteristics (Shapiro and Kostanich (1988)). Both the RDD and the area sample are incomplete in the NSAF, but to different extents. The coverage of the frames for these two components of the NSAF and procedures taken to dampen their effects are discussed in the remainder of this section.

There are three different weighting adjustments that are intended to reduce the bias due to undercoverage: (1) an adjustment for truncation of the area sample; (2) poststratification of households; and (3) poststratification of persons. The first adjustment is discussed in this section and the others are discussed in Section 4.

2.1 Telephone Coverage

The RDD part of the sample used a list-assisted method for sample selection. For this survey, list-assisted was defined as restricting the sample to blocks of 100 telephone numbers having one or more listed telephone numbers. Previous research (Brick *et al.* 1995) estimated that only about 2 or 3 percent of telephone numbers are missed with this procedure. Furthermore, the Brick *et al.* study indicated that there seems to be only slight differences between the social and economic characteristics of the covered and omitted household. Consequently, it was not considered necessary to make special adjustments for the undercoverage in the RDD component.

2.2 Nontelephone Coverage

A special adjustment before poststratification will partially compensate for the undercoverage of nontelephone households from the area sample; the percentage excluded is larger in the nontelephone than the RDD component. As shown in Table 1, about 7 to 10 percent of nontelephone households were in the excluded block groups in the 1990 Decennial Census data file, but we expected to find a somewhat higher percentage of nontelephone households in the excluded block groups in 1997. The differences turned out to be much greater than we expected. It is likely that the high omission of nontelephone households in 1997 is greater than reported in the 1990 data due to a regression to the mean phenomenon. Under this hypothesis, there is some drifting toward the average for both included and excluded block groups of nontelephone households. Since the block groups excluded from the sampling frame had the lowest reported percentage of nontelephone households in 1990, the percentage of nontelephone households in 1997 in these block groups probably drifted upward in the direction of the overall mean percentage. Below, we present some evidence that suggests this probably accounts for a large part of the shortage of nontelephone households in the area component mentioned earlier.

A special coverage adjustment before poststratification was developed that uses block group characteristics in the adjustment. This more detailed level of geography seems likely to produce adjustments that reduce undercoverage bias. It also provides an opportunity to isolate and measure the effect of the adjustment associated with the planned undercoverage of block groups.

Several methods were considered for the exclusion of block groups from the area sampling frame, most of which can be expressed as an adjustment factor applied to the weights. For example, we could have simply multiplied the weights of the responding nontelephone households by the ratio of the total number of block groups in the population to the number of block groups in the sampling frame. This method was rejected because it treats all block groups equally and, by design, the percent of nontelephone households in the excluded block groups is lower than in the included block groups (about 55% of the block groups were excluded while only 8% of the nontelephone households in 1990 were excluded). This method of adjusting would have overcompensated for the undercoverage and estimated many more nontelephone households than actually exist.

A model with more appeal posits that a nontelephone household from the covered frame is exchangeable with a nontelephone household in the noncovered frame. This model can be implemented as an adjustment factor that is the ratio of number of nontelephone households in the population to the number of nontelephone households in block groups in the sampling frame. Both of these counts can be obtained from the 1990 Census tabulations. The ratio is

$$R_{1(c)} = \frac{T_{\cdot(c)}}{T_{f(c)}} \qquad (1)$$

where $T_{f(c)}$ is the number of nontelephone households in the sampling frame in class c, and $T_{\cdot(c)}$ is the number of nontelephone households in the entire population in that class and site. An initial analysis showed that percent in poverty in each block group is a good variable to use to form adjustment classes.

Another undercoverage adjustment accounts for sampling error in the particular sample of block groups selected for the NSAF sample. This ratio is

$$R_2 = \frac{T_f}{\hat{T}_f} \qquad (2)$$

where the denominator is the estimated number of nontelephone households from the 1990 Census data using the NSAF sample of block groups for a site. Note that the sample size in the area sample is too small to do this adjustment at a fine level of detail, so only the state level adjustment is used.

A weakness in the model as formulated thus far is that it is a static model; the model assumes that the composition of nontelephone households in 1990 is identical to that in 1997. The regression to the mean analysis suggests this is not valid. To address this deficiency, the model can be further extended to account for changes between 1990 and 1997 by assuming that one uncovered nontelephone household from 1990 can be exchanged with *r* uncovered nontelephone households in 1997. If the regression to the mean hypothesis is correct, then *r* should be greater than one.

We arranged for the Census Bureau to prepare special tabulations from the 1997 Current Population Survey (CPS) that, when combined with estimates from the 1990 Census files, can be used to estimate an improved weight adjustment, referred to as *r*. The Census Bureau produced 1997 CPS estimates of the percent of nontelephone households in the block groups in the sampling frame and overall. The estimate of *r* for a site is

$$r = \frac{\dfrac{T_f}{T.}}{\dfrac{\hat{T}_{f.97}}{\hat{T}_{..97}}} \qquad (3)$$

where the denominator of *r* is the percent of nontelephone households in the block groups in the sampling frame based on the 1997 CPS estimates. For example, suppose that in a particular state 92 percent of the nontelephone households were in the sampling frame in 1990, but the CPS estimate for this percent is only 88 percent for 1997. The value of *r* in this case would be 1.045 (92/88). Since the CPS estimates of the percent in the sampling frame are based on a limited sample size, *r* is only estimated at the overall site level.

Table 2 shows some preliminary calculations of the value of *r*. The large adjustment ratios for the nontelephone households in the last column of the table support the regression to the mean hypothesis. This is especially compelling since the ratio for all households is close to unity, suggesting the difference is really due to the transitory nature of nontelephone status in households. Further analyses of these data are planned to examine the differences between the covered and uncovered populations for the NSAF.

The area undercoverage adjustment is then the product of the three factors, $R_{1_{(c)}}$, R_2, and *r*. As a part of a later analysis, the effect of several potential undercoverage adjustment factors such as this product will be examined after the survey weighting is completed.

Table 2
Preliminary computations of percent excluded
based on the 1997 CPS

State	All households			Nontelephone households		
	Percent included			Percent included		
	Expected	CPS '97	Ratio	Expected	CPS '97	Ratio
Alabama	59.1	58.4	1.0	92.7	83.9	1.1
Califomia	41.0	40.6	1.0	92.7	63.3	1.5
Colorado	42.1	41.6	1.0	91.2	61.5	1.5
Florida	45.4	43.3	1.0	90.9	69.4	1.3
Massachusetts	29.6	30.7	1.0	91.0	49.5	1.8
Michigan	40.1	38.2	1.0	90.2	70.8	1.3
Minnesota	39.3	38.6	1.0	90.9	64.3	1.4
Mississippi	64.1	62.7	1.0	90.3	71.7	1.3
New Jersey	33.2	32.1	1.0	94.4	64.2	1.5
New York	41.3	43.2	1.0	92.5	72.4	1.3
Texas	54.7	54.4	1.0	92.3	71.0	1.3
Washington	46.3	49.0	0.9	93.9	69.8	1.3
Wisconsin	40.5	43.1	0.9	90.8	90.4	1.0
Balance, U.S.	57.1	50.2	1.1	92.0	72.3	1.3

3. RESPONSE RATE ADJUSTMENTS

The second issue that requires special attention in the NSAF is the effect of nonresponse on the survey estimates. Nonresponse bias may be serious if the response rate is low and the characteristics of the respondents and nonrespondents are very different. Below, we discuss the general approach to adjusting for nonresponse in the telephone and area samples at the household level. Person level nonresponse adjustments are used in the survey but are not discussed here.

3.1 Telephone Response Rates

RDD surveys generally produce higher levels of nonresponse than surveys conducted in-person. Massey *et al.* (forthcoming) review RDD surveys conducted for official purposes since 1990 and their findings indicate that it is very difficult to achieve a response rate of 60 percent in these surveys. With this level of response, nonresponse adjustments are very important because they tend to reduce response bias. These results apply directly to the NSAF RDD component.

One of the main problems in introducing nonresponse adjustments in RDD surveys is that it is very difficult to obtain information on characteristics of nonrespondents, other than geography, that are adequate for refined adjustments. One variable that is not often considered in RDD surveys but should be useful is whether or not the telephone number is listed. Listed numbers are expected to have higher response rates because it is possible to target these households for more intensive efforts.

We examined this possibility in the survey itself. First, we included several experiments to test the effectiveness of several plausible methods of improving response. The results of these experiments will be reported elsewhere, but two important findings related to nonresponse adjustments are summarized here. In one experiment, a $25 incentive for completion of the screener was offered as part of refusal conversion in a follow-up telephone call. The procedure was ineffective. The conversion rates were virtually the same for households offered the incentive and those that were not. In the second experiment, a federal-express letter, with or without a $5 payment inserted in the letter, was mailed to initial refusals before the follow-up telephone calls. Both the federal express mailing and the incentives were very effective and appeared to be additive. Of course, these mailings are possible only to the somewhat over 50 percent of the sample for whom we can locate a mailing address.

The mailing to listed households produced an impressive increase in the response rate. Thus, these types of measures (in addition to any differences that occur because the households are listed in the first place) resulted in an important difference in response rates between listed and unlisted numbers. (The incentive included as part of the interview, which could be applied to both listed and unlisted households, did not increase response rates.) These findings emphasizes the usefulness of separate weighting nonresponse adjustments for these two subsets of the sample.

Nonresponse rates generally vary according to other variables such as metropolitan status (MSA) status. The Donnelley Company can provide information on census-type variables for telephone exchanges. We are examining such variables as percent of population that is black or Hispanic, percent of households in specified income ranges, percent renters, percent children or college graduates which could be useful in establishing weighting cells. These variables will be used to create adjustment cells for standard nonresponse adjustments.

Another feature of RDD surveys is that the adjustments require determination of the response rates in the various cells, and this has some elements of uncertainty. For example, only some of the ring-no-answers and answering machines are residential. We have used information from other studies to estimate the number of residential nonrespondents, but the application of these results to individual adjustment cells introduces some uncertainty.

3.2 Nontelephone Response Rates

The area sample also suffers from nonresponse, but the response rate is much higher in this component. The response rate for households has two stages: (1) nonresponse arising when households in the sampled segments were screened to determine whether they were nontelephone; and, (2) nontelephone households which refused to complete the household interview. The response rate to the first stage is very high and the response rate to the second stage is also high, as compared to the RDD response rate. Thus, response bias is likely to be less in the nontelephone component.

Nonresponse adjustment cells will also be used for the nontelephone component, with classes at the segment level. Any segments that are small will be combined with others in the same PSU.

4. POSTSTRATIFICATION

As in most household surveys, an important objective of poststratification is to dampen potential residual biases arising from a combination of response errors, sampling frame undercoverage, and nonresponse. Another objective is to reduce sampling errors, a necessary objective in this survey because the NSAF sample sizes within states are fairly modest for some subclasses. In general, the sample will be poststratified to as many independent figures as possible. We use the term poststratified loosely to include raking, a form of multidimensional poststratification.

Five guiding principles determine the poststratification adjustment procedure in the NSAF:

1. Estimates from other surveys are used as controls for poststratification only if they have low bias and significantly lower variances than NSAF estimates.
2. Control totals are well-defined and consistently reported variables such as age, race, sex, and telephone status.
3. Controls at a site or state level are preferred to national controls.
4. Control totals must be current or very stable over time.
5. Decisions about poststratification are made without examining the unadjusted NSAF estimates.

At the household level, poststratification will be used to adjust to the number of telephone and nontelephone households in each state. These two categories will be further split in three groups: households with children, households with only elderly persons, other households. The sources of household control totals are the 1990 Census extrapolated to 1997 using the CPS.

The major source of person level estimates for poststratification are population estimates by age, race/ethnicity, and sex produced by the Census Bureau. Although this information is for 1996 rather than 1997, projections of the data for a year should be quite reliable. CPS estimates of home ownership and educational attainment are also being examined for person level poststratification.

5. CONCLUSION

The NSAF is designed to produce reliable estimates of characteristics related to the changes associated with revisions in social policies affecting the non-aged low-income population for 13 selected states and the nation. Two key statistical issues that must be addressed in the estimation phase are undercoverage and nonresponse bias. Bias due to undercoverage is most serious for the area sample of nontelephone households, while the bias due to nonresponse is most serious in the RDD telephone sample.

The plans for adjusting the basic weights to reduce these two sources of bias are presented in this paper. For nontelephone undercoverage, a special adjustment based on Census data is planned. For nonresponse bias in the RDD sample, a standard weighting class adjustment will be used with listed status, a variable that has rarely been used in RDD surveys. The survey weights will be poststratified to population control totals to reduce bias that remains in the estimates.

In addition to these efforts to reduce the biases, special studies are planned to examine residual noncoverage bias in the area sample and nonresponse bias in the telephone sample. These efforts included followup data collection, analyses of data from other sources, and modeling.

REFERENCES

Brick, J.M., Waksberg, J., Kulp, D., and Starer, A. (1995). Bias in list-assisted telephone samples. *Public Opinion Quarterly*, 59, 218-235.

Massey, J.T., O'Connor, D., and Krotki, K. (forthcoming) (1997). Response rates in random digit dialing (RDD) telephone surveys. *Proceedings of the 1997 Section on Survey Research Methods, American Statistical Association.*

Shapiro, G., and Kostanich, D. (1988). High response error and poor coverage are severely hurting the wave of household survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 443-448.

Waksberg, J., Brick, J.M., Shapiro, G., Flores-Cervantes, I., and Bell, B. (forthcoming) (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the 1997 Section on Survey Research Methods, American Statistical Association.*

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

AN INVESTIGATION INTO THE POSSIBLE USE OF MULTILEVEL MODELS BASED ON SURVEY DATA TO UPDATE CENSUS ESTIMATES FOR SMALL AREAS

P. Heady, V. Ruddock and H. Goldstein[1]

ABSTRACT

Data from the British census in 1991 are used by central government to allocate financial resources to small areas. Six years later there is concern that the characteristics of the populations of small areas may have changed and that the resulting allocation of resources is inequitable. One option is to use Structure Preserving Estimation (Purcell and Kish 1980) to update the local census statistics using trends estimated from more recent national survey data. This paper investigates the potential for estimating trends in census statistics using both a survey estimate of a national trend and survey estimates of local trends. A weighted linear combination of an estimate of the national trend and an estimate of the local trend is used to discuss the amount of survey data needed to produce accurate estimates of local trends. A multilevel model is used to estimate the variance of the local trends which is compared with the variance of the observed local trends in the census statistic between 1981 and 1991. Finally we discuss the potential of auxiliary variables to explain between area variation and improve the precision of small area estimates.

KEY WORDS: Multilevel models; Small area estimation; Census and survey data.

1. BACKGROUND

In Great Britain, as in many other countries, central government allocates some financial resources between local authorities on the basis of census statistics. However, the British census is only carried out once every ten years (in 1981, 1991, 2001 *etc.*), and during those 10 years the relevant characteristics of local populations may change. There is concern that, as the intercensal period progresses, the census figures may become progressively less adequate indicators of the current situation of local authorities – and that, as result, financial allocations based on census statistics may become progressively less fair. We therefore need to find ways of updating census statistics during the intercensal period.

How we might do so depends on the information we have available and the assumptions that we are willing to make. Simplifying slightly, it is possible to identify three broad approaches.

The first approach resembles the Structure Preserving Estimation proposed by Purcell and Kish (1980) in that we have information – provided by a survey – about how a marginal distribution has changed (in this case the national proportion of individuals or households in certain categories) and distribute this change between local authorities by making the assumption that their relative position – according to some appropriate definition – has not changed. Equivalently, one can say that a standard up rating factor – defined on some appropriate scale – is applied to the results for each authority.

The second approach supplements these data with survey data for each local authority. The optimum estimate for a particular authority is then a weighted combination of the estimate that assumes a standard up rating factor, and the estimate derived from the data collected in the authority itself. The choice of optimum weighting depends on the relative magnitude of between-authority and within-authority variance – and therefore leads naturally to a formulation in terms of multilevel modelling (see Ghosh and Rao for a review of the range of methods available). In multilevel modelling terms, the variability of local trends around the national average is seen as a random effect.

In the third approach, auxiliary variables are introduced to help predict the local trends. The difference between local and national trends is no longer assumed to be purely random – but is ascribed in part to fixed effects associated with relevant covariates. These might include related variables for which administrative time series are available. They might also include local statistics taken at the time of the last census, if these define area-types in which the subsequent trends turned out to be different. If fixed effects of either kind can be identified, they allow us to reduce the amount of random variability associated with our estimates.

Important though the third approach is, in this paper we restrict our quantitative analysis to the first two. We will evaluate the performance of these approaches in estimating local trends in a particular census statistic, the proportion of households containing only one adult, between the 1981 and 1991 censuses. The criterion we will use is the estimated average weighted mean square error. This is defined as follows. For any estimator of local trends $\hat{\beta}_i^{any}$,

[1] Patrick Heady and Vera Ruddock, Methods and Quality Division, Office for National Statistics, 1 Drummond Gate, London, SW1 V2QQ, United Kingdom, and Harvey Goldstein, Department of Mathematics, Statistics and Computing, Institute of Education, 20 Bedford Way, London, WC1H 0AL, United Kingdom.

$$\text{average weighted } \operatorname{MSE}(\hat{\beta}_i^{\text{any}}) = \sum_i k_i \operatorname{MSE}(\hat{\beta}_i^{\text{any}})$$

$$= \sum_i k_i \underset{s_i \in U_i}{E} \left(\hat{\beta}_i^{\text{any}} - \beta_i\right)^2 \qquad (1)$$

where k_i is the weight for each local authority, proportionate to the number of households in the local authority in the 1981 census scaled so the weights for all 127 authorities sum to 1.

2. THE DATA AND SOME BASIC ESTIMATES

A scatterplot of the proportion of households containing only one adult as measured by the 1991 census versus the proportion for 1981 (Figure 1) showed that this proportion increased in all local authorities over the period 1981-1991. There was also some change in the relative position of different local authorities, although this variation in the amount of increase was small relative to the overall range of values in either census year – and relative to the overall national trend.

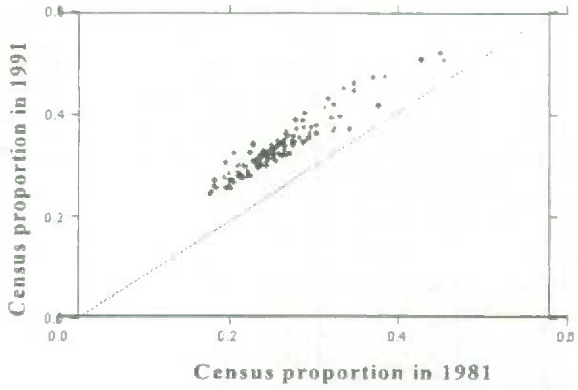


Figure 1. Relationship between the proportion of households containing only one adult in a local authority in 1981 and 1991.

Since we planned to estimate local trends using logistic models we calculated the true annual trend as one tenth of the difference in the logics of the proportions in 1981 and 1991, and calculated local trends in the same way. The national trend was 0.036, the variance was 0.000055 – equivalent to a standard deviation of 0.0074. In the rest of this paper, when we discuss trends or apply the weighted mean square error criterion, the trends in question will refer to the logics – not to the proportions themselves. We restrict our attention to the average local trend for a local authority for the whole period 1981-1991, though in reality we believe the true local trends would probably have varied during that time.

It would be convenient if we had survey data to cover the whole 1981-91 intercensal period, but unfortunately the earliest useable desegregated Labour Force Survey (LFS) data set only goes back to 1985. However we can supplement this with survey data collected after 1991, so that our full survey data set covers the period from 1985 to 1994. It simplifies the modelling process if the survey data are unclustered within each local authority – and so we have restricted this paper to the 127 urban local authorities in which the LFS sample was unclustered for all the years between 1985 and 1994[2]. For the years 1985-1991 the data were obtained from household interviews between March and May whereas for 1992-1994 the data included interviews carried out throughout the year. The mean number of households per year in each local authority was 165, but in some local authorities the achieved sample had less than 100 households per year.

We can use this survey data on its own – both to estimate both the national trend and the local trends for each of the local authorities.

We use survey data to estimate $\bar{\beta}^{\text{surv}}$ – the average trend between 1985 and 1994 – by fitting a model of the form

$$\log_e\left(\pi_{t+1985,i} \text{ or } 1 - \pi_{t+1985,i}\right) = \alpha_i + \bar{\beta}^{\text{surv}} t$$
$$y_{t+1985,i} \sim \operatorname{Binomial}(n_{t+1985,i}, \pi_{t+1985,i}) \qquad (2)$$

where

$n_{t+1985,i}$ is the number of households in our survey sample in local authority i at time $t + 1985$.

$y_{t+1985,i}$ is the number of single adult households in our survey sample in local authority i at time $t + 1985$.

Since we used survey data from 1985-1994 to estimate trends over the period 1981-1991 we knew in advance that our survey based estimate of the national trend would be biassed. The survey based estimate of the national trend was

0.0384, with an estimated standard error of 0.0016 (*i.e.*, variance 0.0000026). The difference between the survey based estimate of the national trend and the national trend as measured by the census was 0.0024 – which is not statistically significant. Considering that the two estimates refer to different time periods, the two estimates are amazingly close. We conclude that there is no reason to suppose that survey and census data taken over the same period would provide different estimates of the national trend – apart from the (small) effect of sampling error on the national survey estimate.

3. COMPARING DIFFERENT ESTIMATORS OF LOCAL TRENDS

3.1 A SPREE Estimator

Since the SPREE estjmator of the annual trend in every authority is effectively $\bar{\beta}^{\text{surv}}$, the MSE of the trend for local authority i is $(\bar{\beta}^{\text{surv}} - \beta_i)^2$. Averaging nationally, and relying on the fact that sampling variation is independent of the varying values of the areas themselves, we get

[2] All the analyses and figures in this paper are restricted to data from these 127 local authorities.

$$E(\text{MSE}) = \sum_i k_i E(\hat{\bar{\beta}}^{\text{surv}} - \beta_i)^2 = \sum_i k_i E(\hat{\bar{\beta}} - \bar{\beta})^2 +$$

$$\sum_i k_i (\bar{\beta} - \beta_i)^2 = 0.000055 + 0.0000026 = 0.000058.$$

$\sqrt{E(\text{MSE})}$, the corresponding indicator of expected deviation, is therefore 0.0076.

3.2 An Estimator Based on Local Survey Data Alone

At the opposite extreme from an approach which looks only at national-level data, is one which relies entirely on survey data for the local authority concerned. To estimate the local trends for each local authority, we fitted a fixed effects model in SAS (SAS Institute) which allowed the annual trend to be different in different local authorities, *i.e.*,

$$\log_e\left(\frac{\pi_{t+1985,i}}{1-\pi_{t+1985,i}}\right) = \alpha_i + t\,\hat{\beta}_i^{\text{surv}} \tag{3}$$

$$y_{t+1985,i} \sim \text{Binomial}\,(n_{t+1985,i},\ \pi_{t+1985,i})$$

where $y_{t,i}$ = number of single adult households in local authority *i* in the survey sample at time *t;*

$n_{t,i}$ = number of households in local authority *i* in the survey sample at time *t.*

The average variance of the local trend estimators $\hat{\beta}_i^{\text{surv}}$ was estimated as 0.000337, over six times the variance of the true local trends about the national trend. This is equivalent to a standard error of 0.018, giving an average coefficient of variation of about 0.5 – which is clearly unacceptable. As would be expected from such an unstable estimator, the correlation between the estimated and true values of the local trends – $\hat{\beta}_i^{\text{surv}}$ and β_i – was very small: $r = 0.23$. This is low, even allowing for the fact that the estimate $\hat{\beta}_i^{\text{surv}}$ and the actual β_i refer to different periods.

3.3 An Estimator that Combines National and Local Survey Data

We clearly do not have enough survey data to accurately estimate trends for each local authority using data from that authority alone. However we can incorporate estimates of local trends into an estimator,

$$\hat{\beta}_i^{\text{comb}} = w\hat{\bar{\beta}}^{\text{surv}} + (1-w)\hat{\beta}_i^{\text{surv}},$$

which is a weighted linear combination of the survey based estimators of the national trend and the local trends, and has a lower average mean squared error than either.

We choose the value of *w* which minimises the average mean squared error of $\hat{\beta}_i^{\text{comb}}$ *i.e.*, we minimise

$$M = \sum_i k_i \underset{s_i \in U_i}{E} ((w\hat{\bar{\beta}}^{\text{surv}} + (1-w)\,\hat{\beta}_i^{\text{surv}}) - \beta_i)^2 \tag{4}$$

where U_i is the population of all households within local authority i, k_i is the scaled weight and s_i is the sample of households drawn from local authority i.

If we ignore the small covariance between $\hat{\bar{\beta}}^{\text{surv}}$ and any particular $\hat{\beta}_i^{\text{surv}}$, this gives an expression which is simply a particular example of a well-known result

$$w = \frac{\sum_i k_i \underset{s_i \in U_i}{E} (\beta_i - \hat{\beta}_i^{\text{surv}})^2}{\sum_i k_i\ E(\hat{\bar{\beta}}^{\text{surv}} - \beta_i)^2 + \sum_i k_i \underset{s_i \in U_i}{E} (\beta_i - \hat{\beta}_i^{\text{surv}})^2} \tag{5}$$

Substituting in values we already know, we obtain

$$w = \frac{0.000337}{0.000058 + 0.000337} = 0.85$$

i.e., the combined estimator of β_i with minimum average mean squared error is

$$0.85\,\hat{\bar{\beta}}^{\text{surv}} + 0.15\,\hat{\beta}_i^{\text{surv}}.$$

So, in the event, our survey based estimates of local trends has not made much impact on our combined estimator – which is appropriate given the imprecision of the local survey figures.

3.4 Estimating Local Trends Using a Multilevel Model

In practise we would not know the variance of the true local trends, but we can estimate this variance by fitting a multilevel model, using the package MLn. We estimate the local authority district trends β_i using a multilevel model where households are clustered within local authority districts.

$$\log_e\left(\frac{\pi_{t+1985,i}}{1-\pi_{t+1985,i}}\right) = \alpha + u_i + t\,(\bar{\beta} + \varphi_i) \tag{6}$$

$$y_{t,ij} \sim B(1, \pi_{t+1985,i}),$$

$$u_i \sim N(0, \sigma_u^2),\ \varphi_i \sim N(0, \sigma_\varphi^2),\ \text{cov}(u_i,\ \varphi_i) = 0$$

where *i* stands for local authority and *j* for household.

In our previous notation β_i is estimated by $\hat{\beta}_i^{\text{mult}} = \hat{\bar{\beta}} + \hat{\varphi}_i^{\text{mult}}$.

The estimated deviation $\hat{\varphi}_i^{\text{mult}}$ of a particular local trend from the estimated national trend $\hat{\bar{\beta}}$ is shrunk towards the estimated national trend, the amount of shrinkage relating to the number of households in the local authority. Therefore local survey estimates based on a large number of households make a greater contribution to the combined estimator $\hat{\beta}_i^{\text{mult}}$ than local survey estimates based on a smaller number of households. This is similar to allowing different values of *w* for different small areas in the combined estimator outlined earlier.

We fitted this 2 level logistic model using the package MLn (Rasbash, Yang, Woodhouse and Goldstein 1995) using Restricted Iterative Generalised Least Squares estimation and a second order Taylor expansion of the non linear term about the level 2 (local authority) residuals (Goldstein and Rasbash 1996).

The correlation between the estimated $\hat{\beta}_i^{\text{mult}}$ and the true values β_i is still small, $r = 0.15$[3] (Figure 2).

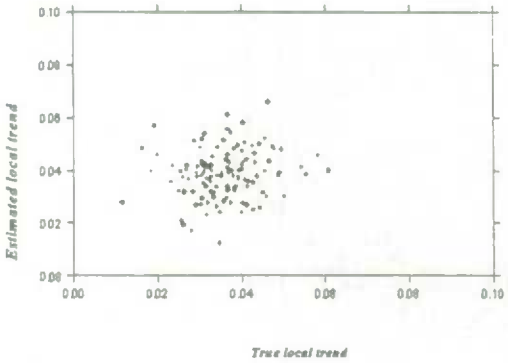


Figure 2. Relationship between survey trends estimated from survey data using MLn and true local trends.

Table 1 shows the estimated fixed and random effects from our model.

Table 1

Fixed effects			
Parameter		Estimate	Standard Error
α		-0.9529	0.0236
$\bar{\beta}^{\text{mult}}$		0.0387	0.0021
Random effects			
Parameter	Level	Estimate	Standard Error
σ_u^2	2	0.0588	0.0082
σ_φ^2	2	0.0002	0.000065
var / bin - var	1	0.9991	0.0031

There is considerable clustering of single adult households within local authority districts as measured by the estimate of σ_u^2 – which is consistent with the wide range of proportions for local authorities shown in Figure 1. The estimated variance of the deviations of the local trends φ_i from the national trend is significantly different from zero (by about three times its own standard error). Unfortunately this estimated variance (0.0002) is higher than the true value of 0.000055 obtained from the census data. One explanation for this might be that the variance of the true local trends over the period 1985-1994 was greater than the variance of the local trends over the period 1981-1991. It is also possible that differences between survey and census data collection techniques mean that their trend estimates are subject to different biases at the local level – over and above any national level differences. These issues would need to be resolved before we went ahead with local trend estimates based on multi-level models, but we would not expect them to provide fundamental difficulties if the more serious problems due to limited survey sample sizes could be resolved.

4. STRATEGIES FOR IMPROVING SURVEY – BASED ESTIMATES OF LOCAL TRENDS

The small contribution of local survey data to our combined estimator of local trends suggests that our data as it stands is not sufficient to produce estimates of local trends which are accurate enough to improve on a SPREE type estimator of local proportions. We conclude this paper by considering two alternative strategies for seeking to increase the contribution of local data to survey based estimates of local trends.

4.1 Increasing the Contribution of Survey Data to Estimates of Local Trends

Our survey data set was not large enough to have a large impact on our combined estimates of local trends. Since the contribution of the survey data is calculated from the estimated variances of the $\hat{\beta}_i^{\text{surv}}$, and these variances are proportional to the inverse of the sample size in a particular local authority we can examine the impact of increasing the survey sample size on the expected value of the average mean squared error of the combined estimator[4]. In particular how much larger would the survey data need to be to reduce the expected value of the average mean squared error of the combined estimator by fifty percent? To obtain this reduction we have to increase the sample size of the survey by a factor *f*, such that

$$f = \frac{2ab - bc}{ac},$$

where *a* is the variance of the true local trends about the national trend, *b* is the sampling variance of the local trends estimated from the survey data and *c* is the original expected value of the average weighted mean squared error. For our Labour Force Survey data we would need to multiply the achieved sample size by 8.2 to obtain a fifty percent reduction in the average mean squared error of the combined estimator. With this enlarged sample the weight *w* would be 0.43 *i.e.*, a greater weight would be given to the local area estimators than the national estimator.

4.2 Using Appropriate Auxiliary Variables to Model Differences Between Local Trends

A more feasible approach involves incorporating auxiliary variables into a survey based estimator of the national trend. The differences in the values of these variables between the local authorities will absorb some of the variation between the local authority trends. Since the coefficients of these variables will be estimated using the

[3] The difference between this correlation and the correlation between the raw local survey trends and true values ($r = 0.23$) is due to the differential shrinkage of the estimated deviations of the local survey trends from the national trend in the multilevel model.

[4] This follows by applying a well known result on the properties of the inverse of a matrix (Healy 1986) to the matrix of second partial derivatives of the log likelihood for our logistic model (Hosmer and Lemeshow 1989).

whole data set their estimated standard error will be small and there will be only a small increase in the average mean squared error of the estimator.

We identified two possible auxiliary variables for the proportion of households containing only one adult, but in neither case were data easily available. The first possibility is to use data from local authority records on Council Tax payments. In the UK all households pay a council tax to their local authority. Households containing only one adult can claim a reduction on their annual bill so there must be a record of the proportion of households in each local authority where this reduction applies which could be used as an auxiliary variable.

Another possible source of auxiliary data is the Electoral Register. Each local authority has a register with a record for each individual at an address who is eligible to vote. The proportion of addresses containing only one voter could therefore be used as an auxiliary variable. Although addresses are not quite the same as households, and the register is incomplete, it might nevertheless be a useful auxiliary variable.

If it is important to estimate local trends in census statistics then we need to maintain series of appropriate auxiliary variables recorded at the local authority level. The availability of auxiliary variables has increased in recent years so we should be able to evaluate the gains which may be made using auxiliary variables and survey data to update statistics from the 1991 census when we have data from the UK census in 2001.

REFERENCES

Ghosh, M., and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9, 1, 55-93.

Goldstein, H., and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A*, 159, Part 3, 505-514.

Healy, M.J.R. (1986). *Matrices for Statistics*. Oxford Science Publications, Clarendon Press. Oxford.

Hosmer, D.W., and Lemeshow, S. (1989). *Applied logistic regression*. Wiley Series in Probability and Mathematical Statistics, Wiley.

Purcell, N., and Kish, L. (1980). Postcensal Estimates for Local Areas (or Domains), *International Statistical Review*, 48, 3-18.

Rasbash, J., Yang, M., Woodhouse, G., and Goldstein, H. (1995). *MLn: command reference guide*. London: Institute of Education.

SAS Institute Inc. (1989). *SAS/STAT User's Guide, Version 6. Fourth Edition*, 2, Cary, NC.

SESSION I-11

New Directions in Questionnaire Development, Testing and Design

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

USING DATABASES TO DESIGN, GENERATE AND STORE BUSINESS QUESTIONNAIRES AT STATISTICS CANADA

A.M. Boltwood[1]

ABSTRACT

Statistics Canada is embarking upon an integration of all its business surveys, under tight time constraints. All questionnaires will be based upon harmonised common questions. Each question will be linked to the data requirements of our System of National Accounts. To develop the common questions, customise them for each industry and business, and generate questionnaires, we are using relational databases. This requires us to develop a general typology of business survey questions and their special features.

KEY WORDS: Business survey questionnaire design; Metadata; Database.

1. THE UNIFIED ENTERPRISE SURVEY

Statistics Canada's Project to Improve Provincial Economic Statistics (PIPES) is leading us to take new directions in developing business questionnaires. (See Beelen *et al.* (1997) and Statistics Canada (1997) for more information about PIPES and the UES.). The project offers us the opportunity to improve many survey-taking practises, but under very tight deadlines.

To increase the detail of provincial economic statistics, we are adding new questions, increasing samples and surveying industries not previously covered. To compensate for this increase in paper burden on respondents, we are integrating our existing business surveys, to reduce duplication and track all contact with respondents. Over the next few years, all Statistics Canada's surveys of businesses will become part of the Unified Enterprise Survey (UES).

The UES data collection strategy is to use income tax data wherever possible, and supplement these data with multi-mode questionnaires. A set of "common questionnaires" is being designed to satisfy the new and existing data requirements of our System of National Accounts. This involves harmonising various existing concepts and definitions.

By basing our surveys on common questionnaires, we can improve our ability to combine data from more than one industry for analysis. (See Priest (1995) for more on data integration and medata requirements.). Thus we will consolidate all UESP microdata results in one database. Analysts will require metadata describing all the common and customised questions.

The common questionnaires will be customised for each industry and unit type. ("Unit type" is a combination of three stratification variables: size, complex/simple and enterprise/establishment. See Statistics Canada (1997) for a full explanation.). This could involve adding questions to meet the requirements of external clients, deleting less relevant questions, and sequencing and wording the questions to suit an industry. We wish to maximise the ability to customise by industry, while maintaining the ability to analyse data across industries.

The customised questionnaires will be "personalised" in some cases, using previous responses from the business to reduce paper burden. For example, a firm might only be asked about the commodities that they used last year. (This is already done in Statistics Canada's Annual Survey of Manufactures. See Crysdale (1998) for detail).

"Schedules" will be attached to some questionnaires, to cover cross-economy topics that require a different sampling scheme than the common questionnaires.

We have developed common and customised questionnaires for a Pilot UES, involving seven groups of industries, which will receive questionnaires in March 1998. We are now beginning development of a second, larger group of customised questionnaires and schedules, to be mailed in 1999.

2. CHARACTERISTICS OF OUR BUSINESS QUESTIONNAIRES

The design of a Statistics Canada business questionnaire is quite different from most social questionnaires sent to households. A business survey is more like an income tax form than a friendly conversation:

- the content is similar to financial statements;
- the wording is formal, and requires specialised knowledge of the industry and of accounting principles;
- response is usually required by law;
- there are few skip patterns;
- questions often include columns or grids of cells, to show the breakdown or calculation of a quantity (*e.g.*, sales revenue for a list of commodities);

[1] Alana M. Boltwood, Enterprise Statistics Division, Statistics Canada, 9th floor, Section D6, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6; e-mail: boltala@statcan.ca.

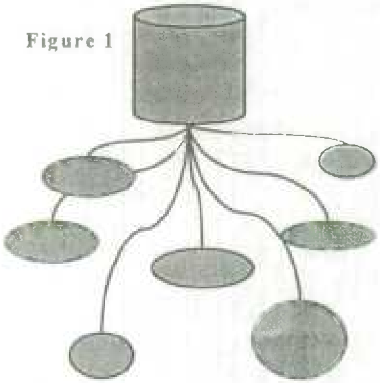
- given the complexity of the content, and the research the respondent must do, we mail paper questionnaires to most respondents; electronic data reporting (EDR) alternatives are sometimes available[2];
- Computer-assisted telephone interviews (CATI) are often used for follow-ups.

3. WHY USE DATABASES IN QUESTIONNAIRE DEVELOPMENT?

Normally, Statistics Canada questionnaires are developed using word processors. The content is reviewed and changed many times before a final version is prepared in desktop publishing and CATI software.

The concept of customising a common questionnaire is illustrated in Figure 1. Each customised questionnaire is represented by an oval, drawn from the canister above. Though the shapes are of different sizes, they are all ovals of the same colour. Thus, if the colour of the canister changes, all the ovals must change colour.

Figure 1



Because of tight deadlines for the UES, the common questionnaires cannot be finalised before we start developing customised versions for each industry. But if the questionnaires are developed in a word processor, it is nearly impossible to change a common question, in a large number of customised questionnaires, quickly and accurately. Thus we are using relational database technology to maintain many customised questionnaires, linked to one common questionnaire.

Creating grid questions requires a complex database structure (see section 5). Entering question text into a structured input screen, rather than a word processor, will draw attention to some potential question design problems. We can also design the database to link questions to data requirements, and thus ensure that every question is justified and every requirement is met.

Once all the questionnaire material is entered in a database, we can generate paper questionnaires that follow a uniform graphic format. The economists who develop questions do not need to spend time formatting questionnaires. We can also output specifications for CATI and EDR application development in a standard format.

4. SEPARATING DATA REQUIREMENTS AND QUESTION DESIGNS

During the development of the Pilot UES questionnaires, we learned that one common data requirement can be met via many different phrasings of a question. For example, the customised questions in Table 1 fit the same basic definition. We use industry-specific terms for independent contractors, and we omit them where not relevant.

Table 1

Industry	Simplified question and definition
Construction:	Number of employees (Exclude subcontractors)
Couriers:	Number of employees (Exclude owner-operators)
Food Services:	Number of employees

We also found that some data requirements cannot be met, because respondents do not have the information, or because we need to use the same definitions as in previous years. In these cases, the customised questionnaire might not include the common question, or it might use a non-standard definition. Our clients need to see a recorded justification for not meeting a data requirement.

We will store the following information, in separate but linked database tables:

- Common data requirements.
- Collection strategy for each requirement, which could be:
  - a tax data element;
  - a common survey question;
  - a derived variable;
  - a justification for not meeting the requirement.
- Text, definition and structure of common survey questions.
- Industry-specific questions (based on requirements which we will not attempt to store centrally).
- Rewordings of common survey questions for specific industries.
- Questionnaire specifications (lists of questions):
  - common questionnaires (templates);
  - customised questionnaires (with additions, deletions, re-orderings, re-wordings);
  - cross-economy schedules.
- Rules for personalisation.

5. A TYPOLOGY OF BUSINESS SURVEY QUESTIONS

To build a database that can create, store and generate survey questions, we must understand precisely their structure. Our observations of many Statistics Canada business questionnaires have helped us develop the following typology of question structures that we wish to make possible in the UES.

Among economic statisticians, the word "question" can refer to a single response cell, a series of single cells on one topic, or an entire page of rows and columns and cells in a

[2] See Hill (1996) for more information on EDR at Statistics Canada.

grid. More precise terms are response cells and groupings of response cells.

A UES questionnaire is made up of one or more sections. Each section contains an ordered series of response cells and/or groupings of response cells. Question text is associated either with an individual cell, or with a grouping of cells.

One response cell corresponds exactly to one data element on a file for processing or analysis. Each space or box on a questionnaire where the respondent can enter a number, a mark or a string of text, is one response cell. Multiple choice questions are an exception: a tick-mark in one of many option boxes can be coded as one categorical data element, so the entire question counts as one "cell" in this typology.

There are various possible types of response cells. Table 1 lists them, along with some display characteristics that must be specified individually for each cell on a questionnaire. In addition to these characteristics, question text must be specified for each cell, and graphical formatting must be specified for each cell type.

Table 2

Response cell type and explanation		Display characteristics
Text	Write-in text strings	Box size
Code	Alphanumeric codes that cannot be used in analytical computations (*e.g.* phone numbers, identification codes, postal codes, dates)	Box format
Numerical	Numbers that can be used in computations	Unit of measure (*e.g.,* $) Scalar factor (*e.g.,*1000s of $) Box size Decimal places
Multiple choice	Respondent selects *one* option from two or more options. Includes yes/no questions. Excludes "check all that apply" questions - see page 5 under *Check-many.*	Option codeset (response categories, which may be grouped under headings, and which may come from a standard classification)
Tick-box	Respondent ticks the box if a statement is true.	*No characteristics to set*
Object	Graphics, audio, video, symbols, computer programs, *etc.*[3] Statistics Canada business questionnaires do not use these currently.	*for example:* Maximum size of object Data transmission method.
Instruction	A paragraph of text with no response box.	*No response box*

Some cell types can be grouped for efficient display on a questionnaire. The type of response cell grouping depends on the cell type. We define three simpler groupings and a generalised grouping type.

Table 3

Grouping type	Cell types that can be grouped
Check-one	Multiple choice
Check-many	Tick-box
Grid	Numerical
Generalized grid	Any cell type or types

Check-one

This grouping type allows us to combine one multiple-choice cell with an "auxiliary cell" such as an Other-Specify text cell, as in Figure 2:

What is your favourite colour?

□ Red

□ Yellow

□ Blue

□ Other - Specify: __________

Figure 2

Check-many

Figure 3 is not a "multiple choice cell" as defined above. Instead it is a series of tick-boxes, the names of which may or may not be drawn from a standard classification. To store responses to this question, "selected" or "not selected" must be recorded for each option in the list, as if each one was a separate tick-box. Check-many groupings can also have auxiliary cells, as illustrated.

What colours do you paint your products?
Check all that apply.

□ Red

□ Yellow

□ Blue

□ Other - Specify: ____________________

Figure 3

Grids

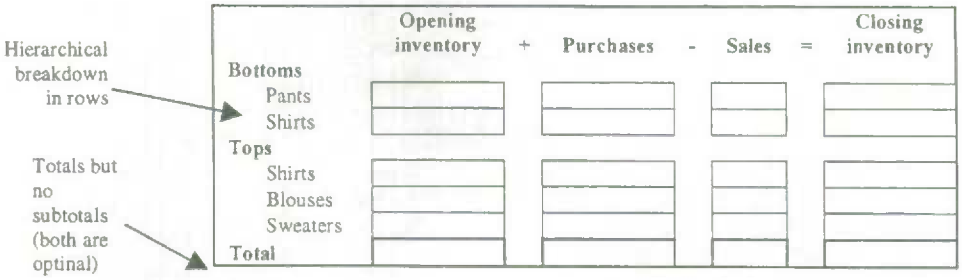
The grid is the most common grouping on Statistics Canada business questionnaires. It groups numerical cells into row, column and "page" dimensions. (A grid might have only one row, one column or one page.) In a well-designed grid, the dimensions represent orthogonal concepts.

Here are some examples of grids. Figure 4 shows that a dimension can be a breakdown of a total, or a calculation. Figures 5 and 6 show that there are many ways to arrange the same cells in a grid; some arrangements are easier for respondents than others. Figure 6 also illustrates page dimensions. In Figure 7, two variables (with different units of measure, in this case) are combined in one grid for display.

Grids can have auxiliary cells of various types. Data elements must be created for each of these cells. Here are examples of all the structures we have observed. In Figure 8, we show various questions added to row labels, including "Other-Specify" lines. Figure 9 shows three different ways respondents can choose the units of measure; the multiple choice and "specify" text cells are auxiliary cells, while the $ and % columns could be considered separate variables.

Grids can also require some special markings and formatting, which do not collect any extra data, so they are not auxiliary cells. These markings are illustrated in Figure 10.

[3] See Fienberg (1997) for more discussion of non-traditional data formats.

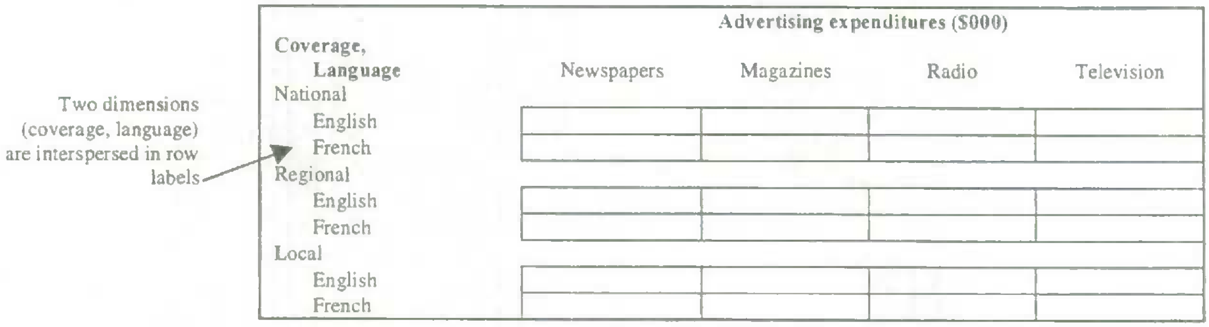


Hierarchical breakdown in rows

Totals but no subtotals (both are optinal)

	Opening inventory	+	Purchases	-	Sales	=	Closing inventory
Bottoms							
Pants							
Shirts							
Tops							
Shirts							
Blouses							
Sweaters							
Total							

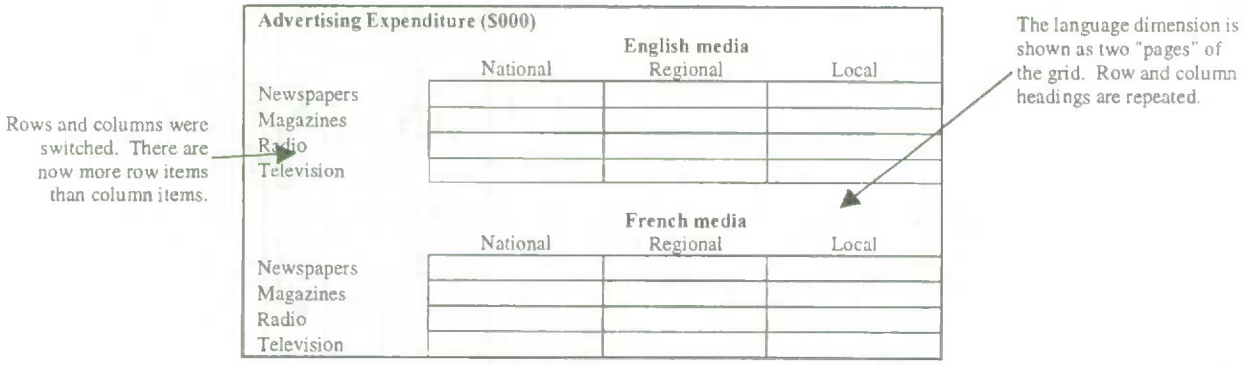
Figure 4



Two dimensions (coverage, language) are interspersed in row labels

	Advertising expenditures ($000)			
Coverage, Language	Newspapers	Magazines	Radio	Television
National				
English				
French				
Regional				
English				
French				
Local				
English				
French				

Figure 5



Rows and columns were switched. There are now more row items than column items.

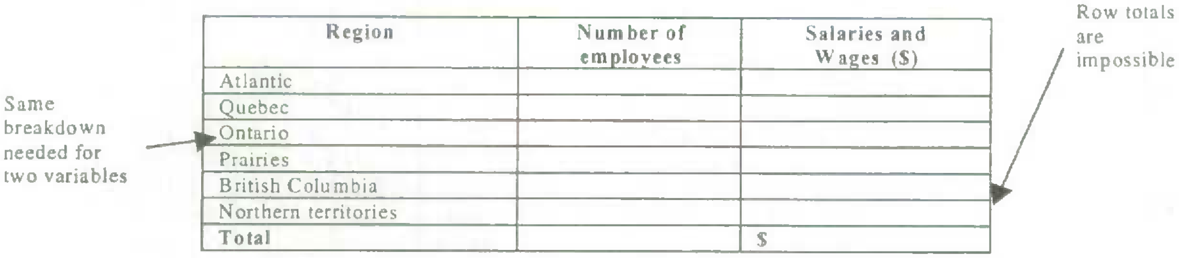
The language dimension is shown as two "pages" of the grid. Row and column headings are repeated.

Advertising Expenditure ($000)

		English media	
	National	Regional	Local
Newspapers			
Magazines			
Radio			
Television			

		French media	
	National	Regional	Local
Newspapers			
Magazines			
Radio			
Television			

Figure 6

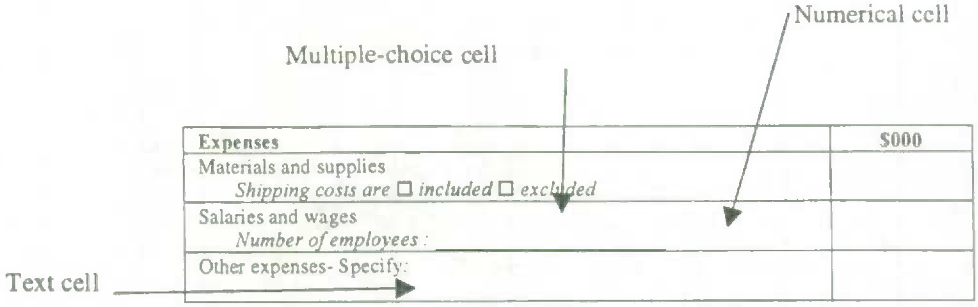


Same breakdown needed for two variables

Row totals are impossible

Region	Number of employees	Salaries and Wages ($)
Atlantic		
Quebec		
Ontario		
Prairies		
British Columbia		
Northern territories		
Total		$

Figure 7

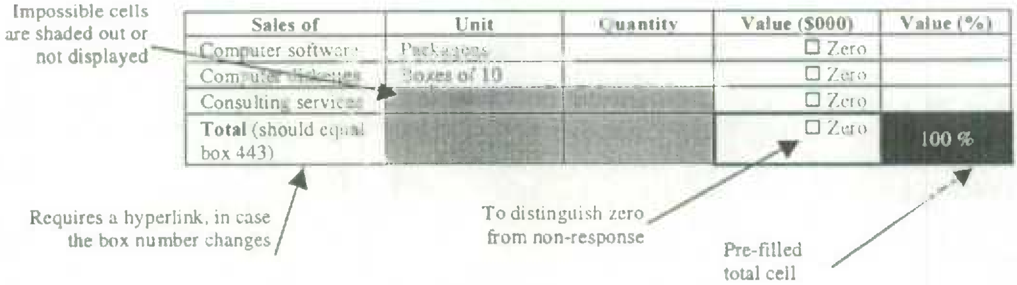


Expenses	$000
Materials and supplies *Shipping costs are* □ *included* □ *excluded*	
Salaries and wages *Number of employees :* ________	
Other expenses- Specify:	

Figure 8

Energy expenditures	Unit	Quantity	Cost ($)	or Cost (%)
Gasoline	□ litres □ gallons			
Natural gas	m^3			
Other- specify type :	*specifyunit*			

Figure 9


Impossible cells are shaded out or not displayed
Requires a hyperlink, in case the box number changes
To distinguish zero from non-response
Pre-filled total cell

Sales of	Unit	Quantity	Value ($000)	Value (%)
Computer software	Packages		□ Zero	
Computer diskettes	Boxes of 10		□ Zero	
Consulting services			□ Zero	
Total (should equal box 443)			□ Zero	100 %

Figure 10

Generalized Grids

A generalised grid allows groupings of any cell type, in rows, columns and pages. This structure could be used to request a list of text strings or codes, to repeat a multiple-choice question, or to create a more complex grid such as Figure 11.

6. SYSTEMS WE ARE BUILDING

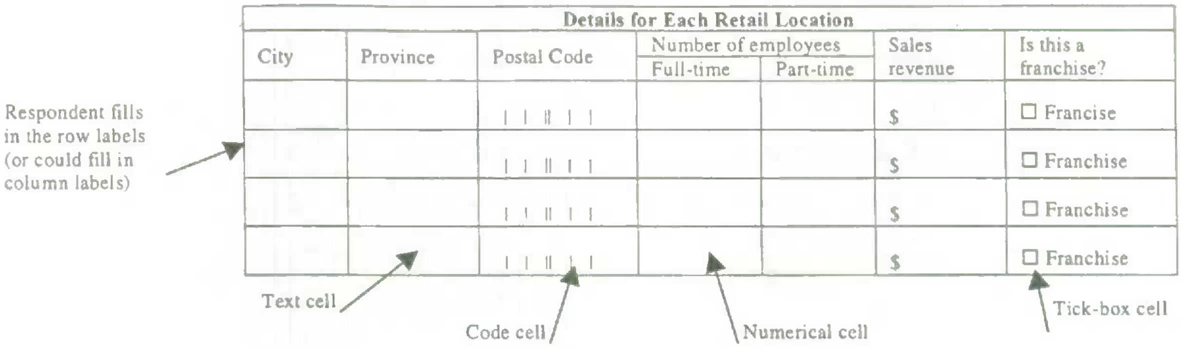
For the Pilot year of the UES, we built a prototype database in MS-Access, called the Question Capture Tool (QCT). Using "rapid application development", we incorporated user needs as they became apparent. The QCT was used to store data requirements and common questions. It also enabled the selection of optional questions for each of seven industry groups.

We found that customization of question wording and ordering was necessary, and the QCT was not designed to allow this. Once the common questionnaire was designed, we halted development and use of the QCT, and used MS-Word to complete the customised questionnaires. (This demonstrated, beyond any doubt, that a word processor would be impractical for updating more than seven customised questionnaires.)

The QCT experience also showed that we needed separate database tables for data requirements, common questions and customised questions. The tool we built could not support the development of separate layers, nor the generation of customised questionnaires, even in draft form.

We are now developing a Content Development Environment which will help us:

- Track our discussions and decisions about harmonising existing surveys: Who said what, when?
- Record and edit data requirements, and their links to questionnaires.



Details for Each Retail Location						
City	Province	Postal Code	Number of employees		Sales revenue	Is this a franchise?
			Full-time	Part-time		
					$	□ Francise
					$	□ Franchise
					$	□ Franchise
					$	□ Franchise

Respondent fills in the row labels (or could fill in column labels)
Text cell
Code cell
Numerical cell
Tick-box cell

Figure 11

- Develop, design and revise common questionnaires.
- Customise and personalise questionnaires for each industry and business.
- Display or print draft questionnaires for internal review and field testing.
- Generate specifications for developing paper forms, CATI and EDR applications (this function is similar to that described in Hunter 1997).
- Measure response burden (*e.g.*: show how many cells are on each customised questionnaire).

We are also developing an Integrated Questionnaire Database to store finalised data requirements, questions and questionnaires. This metadata will be used in survey processing and analysis. Structured documentation will be essential in an integrated processing environment, where many people and systems need to know which business was asked which questions.

ACKNOWLEDGEMENTS

Marcellin Falardeau and Laurie Hill of Statistics Canada have made major contributions to the conception of these databases.

REFERENCES

Beelen, G., Hardy, F., Laniel, N., and D. Royce (1997). Project to improve provincial economic statistics. To appear in *Proceedings: Symposium 97, New Directions in Surveys and Censuses*, Statistics Canada, November 1997.

Crysdale, J.S. (1998). Personalized Questionnaires for Canada's Annual Survey of Manufactures, Business and Trade Statistics. Field Research Paper, Statistics Canada, Ottawa. See also *1997 Proceedings of the Section on Survey Research Methods, American Statistical Association*, Alexandria, Virginia, 1998 (forthcoming).

Fienberg, S.E. (1997). Towards multiple-media survey and census data: rethinking fundamental issues of design and statistical analysis. To appear in *Proceedings: Symposium 97, New Directions in Surveys and Censuses*, Statistics Canada, November 1997.

Hill, L. (1996). Non-sampling error: Can electronic reporting help? In *Proceedings: Symposium 96, Nonsampling Errors*, Statistics Canada.

Hunter, L., and Ladds, J. (1997). Documentation that works. (An integrated approach to survey development, processing and documentation). To appear in Proceedings: *Symposium 97, New Directions in Surveys and Censuses.*

Priest, G. (1995). Data integration: The view from the back of the bus. In *Proceedings: Symposium 95, From Data to Information - Methods and Systems*, Statistics Canada.

Statistics Canada (1997). *Unified Enterprise Statistics Program.* Catalogue no. 68 N0003XPE.

MEASURING THE IMPACT OF WELFARE REFORM: ISSUES IN DESIGNING THE SURVEY OF PROGRAM DYNAMICS QUESTIONNAIRE

J.C. Hess[1] and J.M. Rothgeb

ABSTRACT

The Personal Responsibility and Work Opportunity Reconciliation Act of 1996, more commonly known as the Welfare Reform Act, charged the U.S. Bureau of the Census to evaluate the impact of the law. Toward this end, the Census Bureau developed the Survey of Program Dynamics (SPD). The SPD is composed of two distinct parts: one is an interviewer-administered automated instrument to be answered by an adult respondent and the second is an adolescent self-administered questionnaire. In this paper, we describe challenges we faced in developing, designing and testing these two new survey instruments. Development issues include defining the content of this omnibus survey to meet the needs of the legislation and limiting the scope so as not to overburden respondents and exceed budgetary constraints. Design issues include incorporating both household - and person - level questions to improve the efficiency of collecting data, and administering the questions for the adolescent questionnaire with an audio-cassette player (with headphones) to ensure privacy for the adolescent respondent when answering potentially sensitive questions on various behaviours and practices. Testing issues include conducting cognitive interviews from the paper version of the automated adult questionnaire because of a compressed schedule for pretesting the instrument, and conducting cognitive interviews on an instrument designed to be administered by cassette player with adolescent respondents. We will describe the different challenges we faced and discuss how each was resolved.

KEY WORDS: Welfare reform; Cognitive interviewing; Questionnaire design.

1. INTRODUCTION

President Clinton signed The Personal Responsibility and Work Opportunity Reconciliation Act of 1996, more commonly know as the Welfare Reform Act, on August 22, 1996. One section of the Act charged the U.S. Bureau of the Census:

- To continue to collect data on the 1992 and 1993 panels of the Survey of Income and Program Participation (SIPP) to evaluate the impact of the law on a random national sample of recipients of assistance;
- To pay particular attention to the issues of out-of-wedlock birth, welfare dependency, the beginning and end of welfare spells, and the causes of repeat welfare spells; and,
- To obtain information about the status of children participating in such panels.

Toward this end, the Census Bureau developed the Survey of Program Dynamics (SPD). With current funding, the SPD will extend the 1992/93 SIPP panels through 2001 resulting in 10-years of longitudinal data. This paper describes the challenges we faced in developing, designing, and testing the SPD survey instruments. (See Weinberg *et al.* in this volume for background information about the SPD.)

2. DEVELOPMENT OF THE SURVEY INSTRUMENT

The SPD is comprised of two parts. The first part is called the "core" instrument and includes questions about adults and children. The adult questions, with a few minor exceptions, are asked of all household members ages 15 and over. The core questionnaire was designed for computer-assisted personal interviewing (CAPI.) The second part is a separate self-administered questionnaire (SAQ) for adolescents 12-17 years of age. Provided below is a brief description of the content areas that are included in the various components of the SPD.

2.1 Adult Questions

Employment, Earnings

One of the primary goals of SPD is to chart the economic well-being of families over time in order to evaluate the impact of welfare reform. We collect whether adults are currently working and a detailed account of work-related activities in the past calendar year, including weeks they worked, weeks on layoff, and weeks spent looking for work. We collect detailed employment data for up to four jobs in the previous calendar year.

Income Sources, Amounts, and Eligibility

In addition to earnings, the SPD collects data on a comprehensive list of other income sources, including cash

[1] Jennifer C. Hess, Center for Survey Methods Research/SRD, U.S. Bureau of the Census, Room 3125/4, Washington, D.C. 20233, U.S.A.

and non-cash transfer programs. Data for these items include the type of income received, who received it, months received (if appropriate), and amount received for the previous calendar year. In addition to income, we ask about assets and debts. These items provide information on the economic well-being of persons and households, eligibility for welfare programs, receipt of welfare and other cash and non-cash assistance, and duration of welfare spells.

Educational Enrollment and Work Training

The adult questions on educational enrollment will track the progress of adults toward receiving high school or high school equivalency degrees as well as college and graduate degrees. The work training questions will focus on whether adults received any work training designed to help them look for a job or to train them for a new job. Both educational enrollment and work training are important activities to monitor since participation in these activities is tied to eligibility for receiving welfare benefits in some states.

Disability, Health Care Utilization, and Health Insurance

Questions on disability, health care utilization and health insurance are condensed versions of similar series included as topical modules in SIPP. These questions are included to measure changes in the US health care system and how the changes affect accessibility to government health insurance, such as Medicaid and Medicare, as well as private and employer – provided insurance, utilization of health services, and health-related outcomes.

Food Security

This series of food security questions is a shortened version of the USDA-sponsored Food Security Supplement to the Current Population Survey (CPS) and is intended to measure the subjective experience of hunger. The questions are used as a scale to measure the severity of hunger in a household. Direct changes in the Food Stamp program account for nearly half the total Federal cost savings under the legislation. Additionally, the food stamp benefits of legal immigrants and able-bodied persons ages 18-50 years old without dependents were affected almost immediately by changes in the legislation.

Marital Relationship and Conflict and Adult Depression

Another objective of welfare reform is to encourage marital and family stability. Questions on marital relationship and conflict and adult depression provide indicators of marital happiness and of overall stress that can contribute to marital and family harmony or instability. Changes in program participation and employment can have fairly large and immediate consequences for marital and family stability. These questions are asked of the respondent only and are included following the child-related questions. Because of the personal nature of these questions, the Field Representative will turn the laptop computer toward the respondent and allow the respondent to answer these questions by himself/herself.

2.2 Child-Related Questions

School Enrollment and Enrichment Activities

The school enrollment questions track children's progress through and out of school over time. As household and family conditions change, student progress may be impeded or facilitated. Questions on extracurricular enrichment activities add to the overall portrait of the child's development. The extent to which parents have the financial resources or the time to devote to such activities may be strongly influenced by their program participation and employment in the labor market. These questions provide the basis to study how welfare reforms affect key child outcomes by influencing children's exposure to enriching activities.

Disability and Health Care Utilization

These questions are similar to those asked about adults and are described in Section 2.1.

Child Care

A key objective of welfare reform is to encourage single mothers to enter or re-enter the labor force. Children of these mothers will need to be taken care of during the time the mother is at work. The amount of time children are in child care and the type of care they receive, as well as the stability of care arrangements has been linked with child well-being. If the demand of child care outstrips the supply, or if child care is too costly, there is the possibility that greater numbers of children will be left to care for themselves.

We will ask about all arrangements used since January of the previous calendar year, until the date of the interview. We will know which type of arrangement was used and which months that arrangement was used. This will allow analysts to match the child care data to the employment data for the preceding calendar year. Detailed questions about hours used per week, cost and whether the arrangement is subsidized will be asked only for arrangements currently used.

Child Support and Contact with Absent Parent

Improved enforcement of child support agreements has been highlighted as a cornerstone of welfare reform. Questions on child support will allow researchers to examine the nature of the awards and whether the awards are being followed or enforced. Child support payments are also an important factor in determining the economic status of children living in single-parent families. Another objective of welfare reform is to encourage closer family ties and greater responsibility of parents for their children. Absent parents may participate in and contribute to their children's well-being by providing economic resources or by spending time with them, or both. Questions on contact with absent parents measure the amount of time the non-residential parent spends with their children.

2.3 Adolescent Self-Administered Questionnaire

The Census Bureau, Child Trends, Inc., and the National Institute of Child Health and Human Development's Child

Research Network collaborated to develop the content of the adolescent self-administered questionnaire (SAQ). Adolescence is a time when youths develop the skills and characteristics that increase or decrease the risk of intergenerational dependency. We thought it was important to interview adolescents about their own behaviors because adolescents are often more knowledgeable about their own activities and perceptions than their parents are and collecting data directly from the adolescent will likely improve measurement of these concepts. Provided below is a list of the content areas included in the adolescent questionnaire.

1. Housework and chores; family routines;
2. Parent-child relationships;
3. Parental monitoring;
4. Contact with absent parents;
5. School engagement;
6. Minor problem behaviors and substance abuse;
7. Knowledge of and attitude towards welfare regulations;
8. Dating, early sexual initiation, contraception, and child bearing.

3. QUESTIONNAIRE DESIGN ISSUES

During the development of the SPD, we confronted various questionnaire design issues as the draft questionnaire was reviewed by subject matter and survey methodology experts. Some of the design decisions required major revisions of the questionnaire.

3.1 Adult Questionnaire

The questions on income sources, income amounts, assets and debts were initially designed similar to the March Income Supplement to the CPS. In that survey, all questions about one income source (*e.g.*, who received it and the annual amount received) are asked before asking about the next income source. In order to reduce item nonresponse due to a conditioning effect, we abandoned this design in favor of the design used in the SIPP. In SIPP, we collect an inventory of income sources first. After compiling the different types of income received by all members of the household (*e.g.*, unemployment compensation, Social Security, public assistance), we then ask the amount received from each income source identified. This was done to reduce the likelihood that respondents would stop reporting income sources because they don't want to answer all the follow-up questions.

In a departure from SIPP, which collects this information person-by-person to encourage self reporting, SPD uses household-level screening questions for each income source (*e.g.*, "Did anyone in this household receive any unemployment compensation payments at any time during 1996?") and a single household respondent. This was done to increase efficiency and reduce the amount of time spent collecting this type of information. Increased usage of proxy reporting in SPD will likely lead to some under reporting of income sources. To minimize this type of under reporting, we included extensive use of flashcards during the collection of the income source data. Flashcards listing specific types of income associated with a broader income category are shown to respondents as the Field Representative reads the household-level screening question for that particular income source (*e.g.*, "This is a list of different sources of retirement income. Did anyone in this household receive any pension or retirement income from a previous employer or union, or any other type of retirement income during 1996?"). The theory is that the specific terms included on the flashcard may be more familiar to the respondent than the broad income category included in the question and may improve respondent reporting of income sources, particularly those of other household members. In addition to the use of flashcards, we explicitly encourage respondents to use records to report their earnings and income data.

Another design issue is the collection of income amounts. To assess the impact of time limits associated with welfare reform, we needed to collect both the months a particular income source was received as well as the amount received. In SIPP, which collects data quarterly, amounts are collected for each month of the reference period. In the March CPS Income Supplement, respondents are requested to report an annual amount. Based on research conducted during the redesign of the CPS labor force questions, we opted for a design that allowed respondents to report the income source in the manner that was easiest for them. The computer then calculates an annual amount that the Field Representative confirms with the respondent. This method was shown to reduce item non-response to earnings questions in the CPS.

3.2 Child-Related Questions

As with many demographic household surveys, any household member age 15 or over is eligible to be the household respondent for the SPD; however, for the child-related questions we decided to be more restrictive. Census Bureau experts on children's issues indicated that mothers tend to know more about their children than fathers. They recommended asking the child-related questions of the "designated parent." In the SPD, the designated parent is defined as the mother[2] in two-parent families, the resident parent in single parent families, and as the "person most knowledgeable about the child and his/her activities" in households without a parent. If the mother is not available, we will interview the father. If neither parent is available, we will schedule a call back to talk to the mother. These procedures will increase costs and may also increase item non-response if the Field Representative is unable to collect the data at a later date. However, researchers believe that the benefits associated with improved data quality outweigh the costs and risks.

3.3 Adolescent Self-Administered Questionnaire

The adolescent SAQ contains potentially sensitive questions on problem behaviors, alcohol and drug use, sexual activity and contraception. Protecting the privacy of

[2] "Mother" includes biological, step, and adoptive mothers.

adolescents was paramount in designing this part of the survey. The questionnaire format and procedures mirror those used in the 1992 Youth Behavior Survey (YBS), which asked similar types of questions. Adolescents who are home at the time the Field Representative visits the household will be administered the survey by using an audio-cassette player and will fill out an answer booklet while listening to the tape. The answer booklet contains the answers only and not the questions. Upon completion, the adolescent is instructed to place the answer booklet in the envelope provided and seal it before returning it to the Field Representative. We also developed a separate booklet that contains the survey questions only. This booklet will be shown to parents who request to see the questionnaire. For privacy reasons, the questions are in a different order than those on the tape.

Based on results from the YBS, we estimate that half to two-thirds of the adolescents will not be home at the time of the original interview. We will not make callbacks to administer the adolescent SAQ in person. Instead the Field Representative is instructed to conduct the interview by phone. To protect the privacy of the adolescent during telephone administration, we modified the questionnaire to ensure that answers provided would not reveal the content of the question asked. For example, the following question was included on the cassette tape version of the questionnaire regarding the last time the respondent had sex:

"What method did you or your partner use? Please choose all that apply."

□ No method
□ Birth control pills
□ Condom
□ Diaphragm
□ Foam, jelly or cream
□ Cervical cap
□ Suppository or insert
□ Female condom, vaginal pouch
□ IUD, coil, loop
□ Norplant
□ Depo-Provera, injectables
□ "Morning after" pills
□ Rhythm or safe period
□ Withdrawal, pulling out
□ Other method
□ Not sure

Telephone administration of this question may compromise the adolescent's privacy if he/she answers by giving the name of the method. Therefore, for telephone administration, we modified this question so that the respondent could provide a yes/no response:

> "I'm going to read a list of contraceptive methods. As I read each method, please tell me whether you or your partner used that method the last time you had sexual intercourse."

	Yes	No
Birth control pills	□	□
Condom	□	□
Diaphragm	□	□
Foam, jelly or cream	□	□
Cervical cap	□	□
Suppository or insert	□	□
Female condom, vaginal pouch	□	□
IUD, coil, loop	□	□
Norplant	□	□
Depo-Provera, injectables	□	□
"Morning after" pills	□	□
Rhythm or safe period	□	□
Withdrawal, pulling out	□	□
Other method	□	□
Not sure	□	□

4. QUESTIONNAIRE TESTING

4.1 Adult and Child-Related Questions

The adult and child-related questions in SPD were designed for a CAPI environment. Originally plans called for cognitively testing sections of the automated instrument as they became available. However, we had to work under a compressed time schedule, which meant that cognitive testing and instrument automation occurred simultaneously rather than consecutively. Rather than eliminate the cognitive testing, we decided to cognitively test those sections of the questionnaire that could be conducted, *albeit* somewhat difficultly, on paper. These included all adult- and child-related sections with the exception of the employment, income sources, income amounts, and eligibility questions. The complex skip patterns in these series of questions made conducting an interview on paper impossible.

Testing using a paper instrument proved quite useful. We were able to identify individual questions and series of questions that caused problems for respondents. Problems identified included confusing and unclear reference periods, terms and concepts not well understood by respondents, and items that were too difficult for respondents to answer accurately. Revisions were made to specific items as well as entire series.

4.2 Adolescent SAQ

We conducted cognitive interviews with adolescents ages 12-17 using the version of the SAQ designed to be administered by an audio-cassette player. The objectives of the test included evaluating question understanding, task difficulty, and question sensitivity. To address the first two of these objectives, we conducted interviewer-administered interviews and instructed respondents to "think-aloud" as they answered the questions. Although this method of administration does not mirror the field administration by an audio-cassette player, we believed that using a retrospective technique (with an audio-cassette instrument) would jeopardize our ability to adequately evaluate question understanding and task difficulty.

Three researchers at the Census Bureau's Center for Survey Methods Research conducted the interviews. To ensure comparability across surveys, we developed a protocol beforehand that included additional probing questions to be used at the interviewer's discretion if the respondent did not convey the information while thinking aloud or didn't convey the information after general probes such as, "Could you tell me more about that?" At the end of the protocol we included a few debriefing questions regarding question difficulty and question sensitivity.

Provided below we've identified some of the areas that caused the most problems for adolescents and the revisions that were made to the questionnaire.

1. Respondents tended to ignore reference periods when they were included in the questions, such as "During the past 30 days,...," We revised the questionnaire to include all reference periods in the response options.
2. Respondents tended to interpret lists of examples too narrowly rather than as examples of a broader class of similar activities or events. They would report only about those activities included in the list. We recommended being very cautious of including such lists. In some cases, we deleted the list. In other cases, we revised the list to include items we believed best reflected the concept of interest.
3. Respondents had great difficulty reporting their contact with their absent parent in terms of a "typical month." They tended to report the last time the event happened if it was infrequent, or over report, by guessing, if the event occurred frequently. We revised these questions to ask "how often" the event happens and included categorical response categories ranging from "never" to "everyday or almost everyday."
4. The series of questions on attitudes toward welfare included a response scale ranging from "strongly agree" to "strongly disagree" with a middle category of "I'm in the middle." We found that respondents used this middle category for two purposes: 1) to indicate that they both agreed and disagreed with the statement and 2) to indicate that they didn't know or didn't have an opinion about the statement. We revised these questions to include a specific "don't know" category.
5. The questions on relationship with fathers (in the series on parent-child relationships) referred to the "man who is most like a father to you." A couple of respondents who live in single-parent families with their mothers answered these questions about their mothers since they believe that their mothers are filling the role of both mother and father within their household. We changed the answer category from "there is no one like a father to me" to "I don't live with a biological, adoptive, step, or other father figure."
6. Although there was great concern that the questions about delinquent behaviors, alcohol and drug use, dating and sexual activity would be highly sensitive, only one respondent said he/she was uncomfortable answering one of the sex questions. Some respondents indicated they would be more comfortable using the procedure that will actually be used for the survey (answering the questions privately by listening to a cassette recorder and marking an answer sheet), rather than responding to an interviewer as was done in cognitive testing.
7. The cognitive interviews last from 60 to 90 minutes. We were concerned that the length of the interview and the tedious task of thinking aloud might prove too difficult for adolescents, who are generally portrayed as non-communicative and unable to focus for an extended period of time. Our experience proved contrary to expectations. Adolescents were quite capable of articulating their thoughts in a think-aloud setting and quite able to focus throughout the lengthy interview.[3] Based on our experience, we found a greater need to probe during these interviews than is typically done during cognitive interviews with adult respondents.[4]

5. PLANS FOR EVALUATING THE PRETEST

5.1 Adult and Child-related Questions

We used several methods to evaluate the pretest questionnaire, interviewing materials, manuals, and procedures. Census Bureau staff observed interviews in all four Regional Offices participating in the pretest and completed an Interviewing Observation Form for each household they observe. The form was quite detailed and covered areas of concern such as difficulty administering the adolescent questionnaire at the same time as the adult questionnaire, adolescents' ability to use the audio-cassette recorder to answer the SAQ, flashcard usage, disruptiveness of changing respondents for the child-related questions, and difficulty with specific questions or series of questions (*e.g.*, confusion with the reference period, terms or concepts that were not understood, questions that required extensive probing, *etc.*).

All Field Representatives were requested to tape two complete interviews, with permission of the respondent, to be used for subsequent behavior coding. Behavior coding is the systematic coding of interviewer and respondent interactions. Due to the limited time we have to analyze the pretest data (less than six weeks) and the length of the survey (approximately 60 minutes), coding each question contained in the 90 tapes we hope to get is not possible. We developed a coding scheme that is less systematic and more qualitative in nature than those typically used, but hope that it will still yield sufficient information to identify problematic items and provide some information on possible solutions for fixing the items.

Representatives from CSMR and the Census Bureau's Field Division facilitated debriefings sessions with all Field Representatives participating in the pretest (similar to a focus group). Each session contained 8 to 10 Field Representatives. Topics covered included those contained

[3] Respondents were paid $25 for participating in the research. Admittedly the monetary incentive may affect the respondents' willingness to focus on the tasks at hand.

[4] This observation is based on our limited experience and is not grounded in empirical data.

in the observer form described above, as well as record usage, screen layout of the computerized instrument, problems with the instrument (rostering, demographics, function keys, *etc.*), manuals, training, case management, and the length of the interview. Prior to interviewing, Field Representative's were informed of these debriefing sessions and given a diary to record any problems or observations they have. The diaries were divided into sections based on the content areas outlined in the protocol. Field Representative's were instructed to complete the relevant sections of the diaries on a flow basis so that important information is not forgotten. The diaries helped to keep the discussion focused during the debriefing sessions and make them as productive as possible.

In addition to the evaluation techniques mentioned above, the pretest will also provide data on the length of the survey. We have budgeted for a survey that averages 60 minute per household. We developed a plan to ask selected sections of the questionnaire every two years rather than every year as originally planned and chose sections that we thought would be least likely to change dramatically from year to year (adult and child disability and health care utilization, child enrichment activities, and contact with absent parent). This plan will be implemented if the interview exceeds the targeted 60-minute household average. Subject matter experts preferred this solution because they believed they had already whittled down their series to the bare bones and couldn't adequately measure the concepts with fewer items. The instrument authors preferred this solution, as well, since skipping over an entire self-contained section requires a minimal amount of programming, whereas deleting specific questions may affect skip patterns throughout the instrument and require extensive testing of the revised instrument.

5.2 Adolescent SAQ

We included a series of respondent debriefing questions at the end of the adolescent questionnaire. The adolescent questionnaire contains several series of questions with identical response categories. Owing to concerns about literacy, especially among younger adolescents, we were unsure whether we needed to read all response options for every question. We recorded two different versions of the tape: one in which the answer categories were read for every question, and a second in which the answer categories were read only the first time a series of questions with the same categories was asked. We included debriefing questions at the end of the survey to assess the pace of the tape, whether there was adequate time to mark the answer sheet, and preference for the reading of the answer categories. In addition we asked about privacy concerns if the questions would have been included in the answer booklet (alleviating the need for the audio cassette recorder), the adolescent's ability to concentrate throughout the 30-minute interview, the respondent's level of interest in the survey, and his/her level of comfort answering selected series of potentially sensitive questions.

Child Trends, Inc. will conduct analyses to assess the internal consistency of scales included in the adolescent SAQ, such as those measuring positive relationships with parents, parental monitoring, and school engagement. They will also analyze the frequency of responses such as "don't know," "not applicable," and no response. This will allow us to identify questions that the respondents had trouble understanding or felt uncomfortable answering. In addition, they will examine whether respondents failed to finish filling out the questionnaire, which may indicate that the questionnaire is too lengthy for the respondent's attention span.

6. FUTURE PLANS

Pretest evaluation will be completed in Fall 1997 and a revised draft questionnaire completed by mid-December 1997. Production SPD will be implemented in Spring 1998.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

ASKING QUESTIONS USING COMBINATIONS OF SEQUENTIAL, MATRIX, SINGLE SHEET AND BOOK FORMATS IN A TEST OF THE CANADIAN CENSUS QUESTIONNAIRE

A.R. Gower[1]

ABSTRACT

In planning for the 2001 Census of Canada, innovative changes are being considered for the design of the census form. This paper describes a research study that evaluated the effectiveness of four versions of the census questionnaire: matrix book, matrix sheet, sequential book, and sequential sheet. A total of 90 cognitive, think-aloud interviews took place with respondents who each completed two of the four versions of the census form. Respondents were asked to compare and contrast the two versions of the form that they completed as well as to comment on the respondent-friendliness of each form.

KEY WORDS: Census; Questionnaire testing; Cognitive research; Questionnaire design.

1. INTRODUCTION

The Census of Canada, conducted every five years, collects information that is used for public and private analysis as well as decision-making in many areas concerning the people of Canada. The next census will take place in 2001.

There are two versions of the Census questionnaire. A short questionnaire (*i.e.*, the 2A form) is distributed to 80 percent of Canadian households, while a longer questionnaire (*i.e.*, the 2B form) is distributed to the remaining households. The 2A form collects basic information such as household composition, age, sex, and marital status. The 2B questionnaire collects the basic data together with additional information such as education, ethnic origin, labour force activity, and income. Both questionnaires are respondent-completed.

A number of new directions and changes in the design of the 2A form are being considered for the 2001 Census. Historically, a matrix-type of format has been used in the short census form to ask questions about household members. Consideration is now being given to using a sequential form of questioning for each household member.

Another change involves redesigning the questionnaire to accommodate data capture by scanning, using optical character recognition (OCR) technology. This will necessitate changes to the layout and appearance of the existing form. For example, it seems that a single sheet can be scanned more easily and efficiently than a questionnaire in a multi-page booklet format as has been used in previous censuses. There are also changes in the content of certain questions that are planned for 2001.

During this study, four formats of the 2A census form were examined. The formats were: matrix book, matrix sheet, sequential book, and sequential sheet. The matrix and sequential formats are illustrated in Figures 1 and 2.

2. RESEARCH OBJECTIVES

The main purpose of the research project was to investigate respondent reactions to the four versions of the 2A census questionnaire for the May 1997 Census Test that was conducted in preparation for the 2001 Census.

The study investigated the impact that the design and layout of each of the four versions of the census questionnaire had on respondent behaviour and data quality. The research findings and recommendations were used to ensure that the May 1997 Census Test questionnaires were respondent-friendly with instructions and question formats that could be easily understood and accurately completed.

Cognitive aspects of respondent behaviour such as respondent understanding and the ease of completing the questionnaires were thoroughly explored. The study investigated:

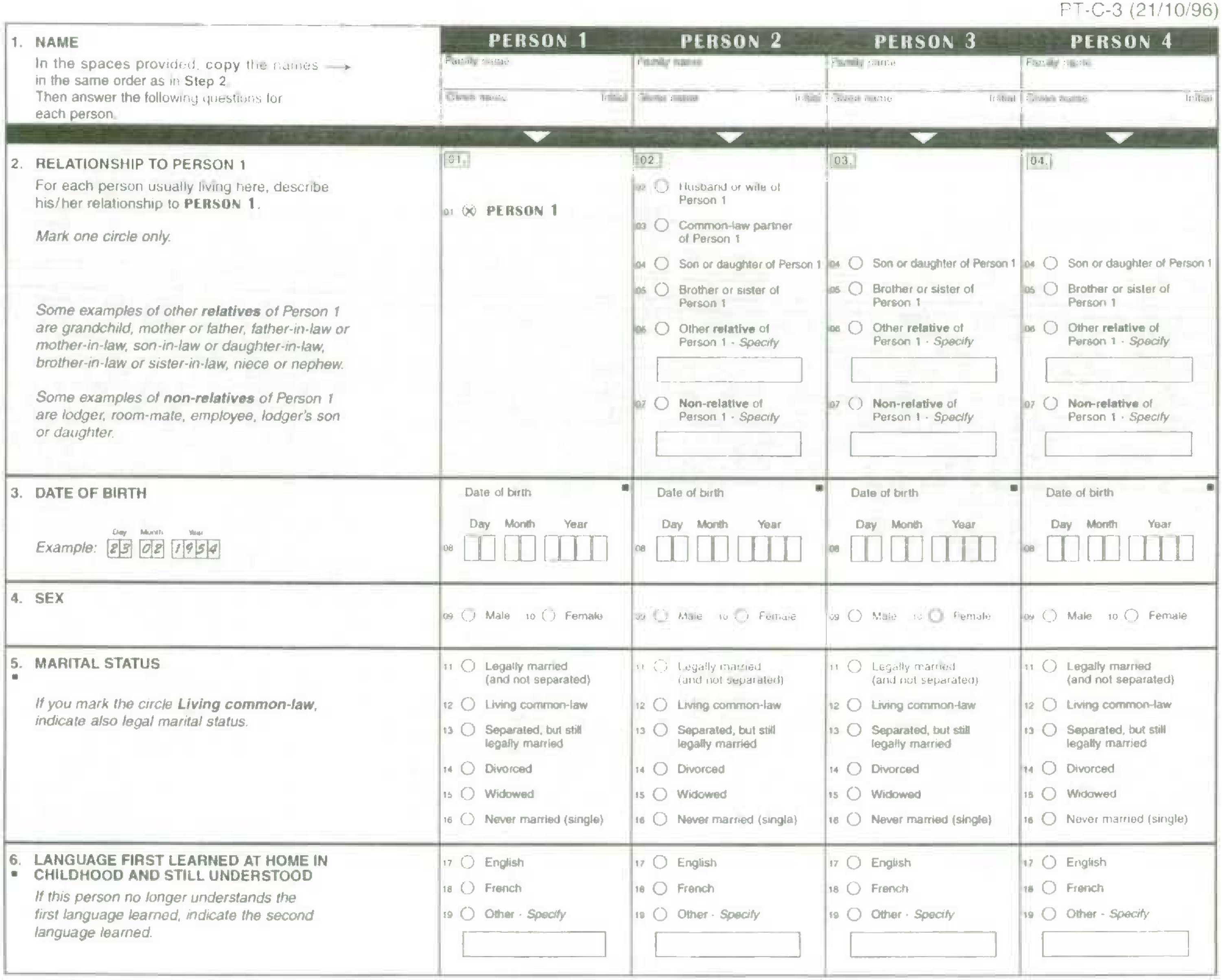
- *how* respondents answered the questions and *why* they answered the way they did;
- problems or confusion that they encountered while completing the census form;
- respondents' understanding of the questions and response categories;
- their reactions to the appearance and layout of the four versions of the census form.

3. METHODOLOGY

Prior to the study, preliminary versions of the census forms were tested with 24 employees of Statistics Canada. The findings of the preliminary testing resulted in a number of revisions being made to the forms before further testing took place.

During the study itself, a total of 90 one-on-one, cognitive (concurrent think-aloud) interviews took place.

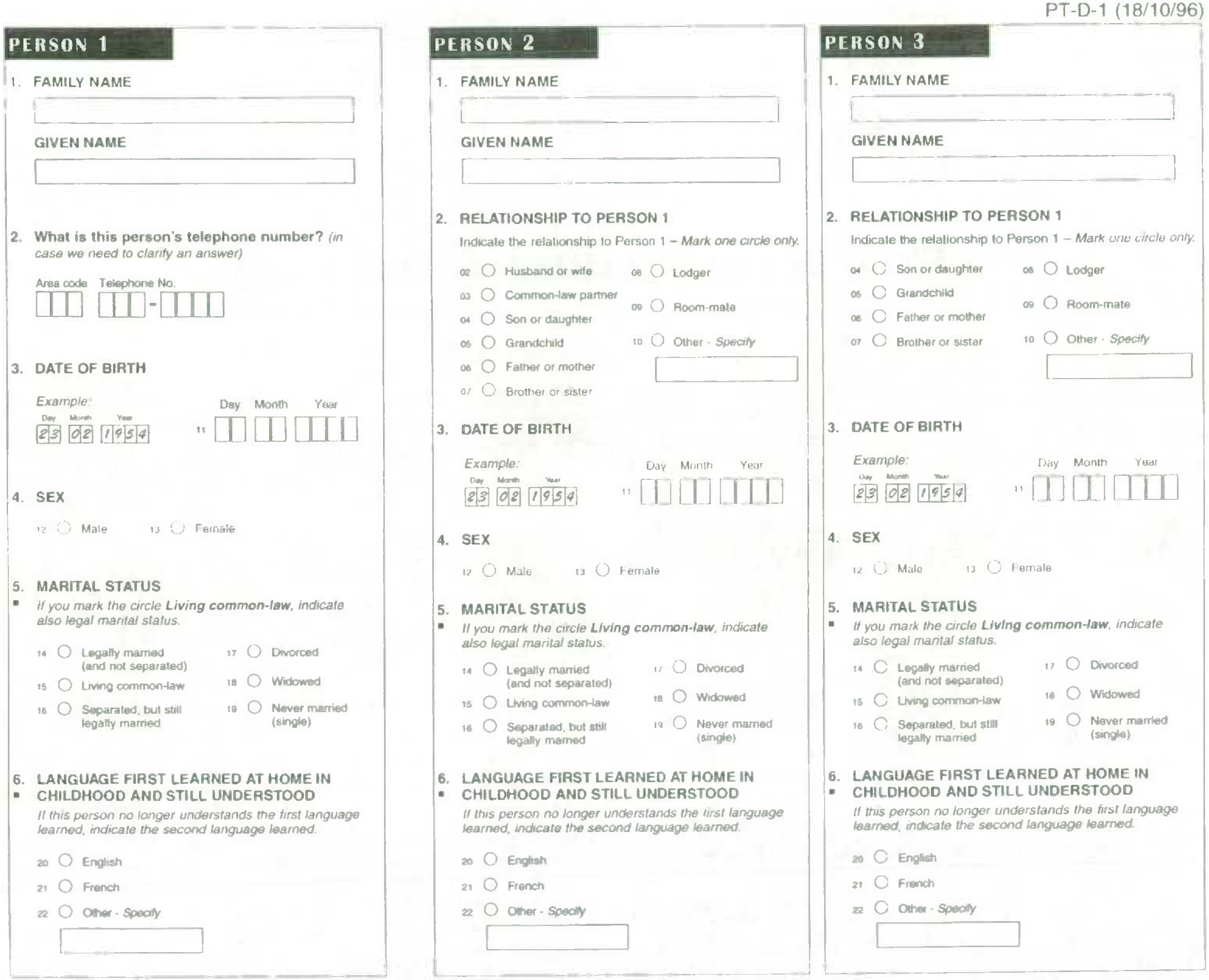
[1] Allen R. Gower, Chief, Questionnaire Design Resource Centre, Statistics Canada, 15-J, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.



PT-C-3 (21/10/96)

	PERSON 1	PERSON 2	PERSON 3	PERSON 4
1. **NAME** In the spaces provided, copy the names → in the same order as in **Step 2**. Then answer the following questions for each person.	Family name Given name Initial	Family name Given name Initial	Family name Given name Initial	Family name Given name Initial
2. **RELATIONSHIP TO PERSON 1** For each person usually living here, describe his/her relationship to **PERSON 1**. *Mark one circle only.* *Some examples of other **relatives** of Person 1 are grandchild, mother or father, father-in-law or mother-in-law, son-in-law or daughter-in-law, brother-in-law or sister-in-law, niece or nephew.* *Some examples of **non-relatives** of Person 1 are lodger, room-mate, employee, lodger's son or daughter.*	01. 01 ☒ **PERSON 1**	02. 02 ○ Husband or wife of Person 1 03 ○ Common-law partner of Person 1 04 ○ Son or daughter of Person 1 05 ○ Brother or sister of Person 1 06 ○ Other **relative** of Person 1 - *Specify* □ 07 ○ **Non-relative** of Person 1 - *Specify* □	03. 04 ○ Son or daughter of Person 1 05 ○ Brother or sister of Person 1 06 ○ Other **relative** of Person 1 - *Specify* □ 07 ○ **Non-relative** of Person 1 - *Specify* □	04. 04 ○ Son or daughter of Person 1 05 ○ Brother or sister of Person 1 06 ○ Other **relative** of Person 1 - *Specify* □ 07 ○ **Non-relative** of Person 1 - *Specify* □
3. **DATE OF BIRTH** *Example:* Day 2 5 Month 0 2 Year 1 9 5 4	Date of birth ■ 08 Day □□ Month □□ Year □□□□	Date of birth ■ 08 Day □□ Month □□ Year □□□□	Date of birth ■ 08 Day □□ Month □□ Year □□□□	Date of birth ■ 08 Day □□ Month □□ Year □□□□
4. **SEX**	09 ○ Male 10 ○ Female	09 ○ Male 10 ○ Female	09 ○ Male 10 ○ Female	09 ○ Male 10 ○ Female
5. **MARITAL STATUS** ■ *If you mark the circle **Living common-law**, indicate also legal marital status.*	11 ○ Legally married (and not separated) 12 ○ Living common-law 13 ○ Separated, but still legally married 14 ○ Divorced 15 ○ Widowed 16 ○ Never married (single)	11 ○ Legally married (and not separated) 12 ○ Living common-law 13 ○ Separated, but still legally married 14 ○ Divorced 15 ○ Widowed 16 ○ Never married (singla)	11 ○ Legally married (and not separated) 12 ○ Living common-law 13 ○ Separated, but still legally married 14 ○ Divorced 15 ○ Widowed 16 ○ Never married (single)	11 ○ Legally married (and not separated) 12 ○ Living common-law 13 ○ Separated, but still legally married 14 ○ Divorced 15 ○ Widowed 16 ○ Never married (single)
6. **LANGUAGE FIRST LEARNED AT HOME IN CHILDHOOD AND STILL UNDERSTOOD** ■ *If this person no longer understands the first language learned, indicate the second language learned.*	17 ○ English 18 ○ French 19 ○ Other - *Specify* □	17 ○ English 18 ○ French 19 ○ Other - *Specify* □	17 ○ English 18 ○ French 19 ○ Other - *Specify* □	17 ○ English 18 ○ French 19 ○ Other - *Specify* □

Figure 1: Matrix format *(62% of original size)*



PT-D-1 (18/10/96)

PERSON 1

1. FAMILY NAME

GIVEN NAME

2. **What is this person's telephone number?** *(in case we need to clarify an answer)*

Area code Telephone No.

□□□ □□□-□□□□

3. DATE OF BIRTH

Example: Day Month Year 23 02 1954

11 Day □□ Month □□ Year □□□□

4. SEX

12 ○ Male 13 ○ Female

5. MARITAL STATUS

▪ *If you mark the circle* ***Living common-law****, indicate also legal marital status.*

- 14 ○ Legally married (and not separated)
- 15 ○ Living common-law
- 16 ○ Separated, but still legally married
- 17 ○ Divorced
- 18 ○ Widowed
- 19 ○ Never married (single)

6. LANGUAGE FIRST LEARNED AT HOME IN CHILDHOOD AND STILL UNDERSTOOD

▪ *If this person no longer understands the first language learned, indicate the second language learned.*

- 20 ○ English
- 21 ○ French
- 22 ○ Other - *Specify*

PERSON 2

1. FAMILY NAME

GIVEN NAME

2. RELATIONSHIP TO PERSON 1

Indicate the relationship to Person 1 – *Mark one circle only.*

- 02 ○ Husband or wife
- 03 ○ Common-law partner
- 04 ○ Son or daughter
- 05 ○ Grandchild
- 06 ○ Father or mother
- 07 ○ Brother or sister
- 08 ○ Lodger
- 09 ○ Room-mate
- 10 ○ Other - *Specify*

3. DATE OF BIRTH

Example: Day Month Year 23 02 1954

11 Day □□ Month □□ Year □□□□

4. SEX

12 ○ Male 13 ○ Female

5. MARITAL STATUS

▪ *If you mark the circle* ***Living common-law****, indicate also legal marital status.*

- 14 ○ Legally married (and not separated)
- 15 ○ Living common-law
- 16 ○ Separated, but still legally married
- 17 ○ Divorced
- 18 ○ Widowed
- 19 ○ Never married (single)

6. LANGUAGE FIRST LEARNED AT HOME IN CHILDHOOD AND STILL UNDERSTOOD

▪ *If this person no longer understands the first language learned, indicate the second language learned.*

- 20 ○ English
- 21 ○ French
- 22 ○ Other - *Specify*

PERSON 3

1. FAMILY NAME

GIVEN NAME

2. RELATIONSHIP TO PERSON 1

Indicate the relationship to Person 1 – *Mark one circle only.*

- 04 ○ Son or daughter
- 05 ○ Grandchild
- 06 ○ Father or mother
- 07 ○ Brother or sister
- 08 ○ Lodger
- 09 ○ Room-mate
- 10 ○ Other - *Specify*

3. DATE OF BIRTH

Example: Day Month Year 23 02 1954

11 Day □□ Month □□ Year □□□□

4. SEX

12 ○ Male 13 ○ Female

5. MARITAL STATUS

▪ *If you mark the circle* ***Living common-law****, indicate also legal marital status.*

- 14 ○ Legally married (and not separated)
- 15 ○ Living common-law
- 16 ○ Separated, but still legally married
- 17 ○ Divorced
- 18 ○ Widowed
- 19 ○ Never married (single)

6. LANGUAGE FIRST LEARNED AT HOME IN CHILDHOOD AND STILL UNDERSTOOD

▪ *If this person no longer understands the first language learned, indicate the second language learned.*

- 20 ○ English
- 21 ○ French
- 22 ○ Other - *Specify*

Figure 2: Sequential format *(62% of original size)*

The Questionnaire Design Resource Centre conducted the research. It took place in Ottawa, Montreal and Toronto during November 1996. The number of cognitive interviews according to location and official language are summarized in Table 1.

Table 1
Number of Cognitive Interviews

Location	English	French	Total
Ottawa	26	17	43
Montreal	-	26	26
Toronto	21	-	21
Total	47	43	90

The respondents were a representative mix of typical respondents, although many were selected according to specific characteristics that were related to the testing of various content items on the census questionnaire. Respondents, for example, included persons from multi-generational households and recent immigrants to Canada.

The study explored respondents' cognitive processes from the time they read the survey questions (or instructions) to the time they wrote their answers. The four distinct processes involved in providing answers (comprehension, retrieval, thinking, and writing the answer) were examined.

Each respondent completed two of the four formats of the census questionnaire. During the interviews, respondents were first asked to complete one version of the census form. They were asked to "think aloud" as they completed the form, to comment on what they were reading or any problems that they were encountering, and to explain how and why they were arriving at their answers. While the respondent was completing the census form and "thinking aloud," the interviewer occasionally probed to determine more details about what the respondent was thinking and the process by which the respondent was completing the questionnaire. After the form had been completed, respondents were asked to provide general comments about their reactions to the form.

Respondents were next asked to complete a second version of the census form, again thinking aloud and providing comments. Both forms were then thoroughly reviewed with the respondents. Finally, they were asked to compare and contrast the two forms, to select which form they preferred, and to provide the reasons for their preference.

Two of the four versions of the questionnaire were randomly assigned to each respondent. All possible orderings and combinations of pairs of forms were completed by respondents as summarized in Table 2.

Because the research was qualitative in nature, findings and conclusions were not representative of the Canadian population who will complete the 2001 Census. They were only representative of the people who participated in the study. However, the results of the testing provided the 2001 Census project team with important insights into respondents' reactions to the proposed forms, and identified potential problems and sources of response error that should be addressed.

Table 2
Order of Presentation of the Census Forms

First version completed	Second version completed
Matrix book	Matrix sheet
Matrix book	Sequential book
Matrix book	Sequential sheet
Matrix sheet	Matrix book
Matrix sheet	Sequential book
Matrix sheet	Sequential sheet
Sequential book	Matrix book
Sequential book	Matrix sheet
Sequential book	Sequential sheet
Sequential sheet	Matrix book
Sequential sheet	Matrix sheet
Sequential sheet	Sequential book

4. FINDINGS

Table 3 summarizes respondents' preferences for each of the four versions of the census form. The table indicates the number of times that each form was completed by respondents, the number of times that each form was preferred (in comparison with the other form that they completed), and the percentage of times that each version of the form was preferred out of the total times it was completed by respondents.

Table 3 indicates that respondents preferred the sequential book format more often than any of the other three formats. The sequential book was preferred 66% of all times that it was completed by respondents, the sequential sheet was preferred 53% of the time, the matrix book was preferred 38% of the time, and the matrix sheet was preferred 34% of the time.

Table 3
Respondents' Preferences for the Four Versions of the Census Form

Format	Times completed	Times preferred	Times preferred as a percentage of times completed
Sequential book	44	29	66%
Sequential sheet	43	23	53%
Matrix book	47	18	38%
Matrix sheet	44	15	34%

In many of these cases, respondents expressed strong preferences when making a choice between two formats. In other cases, respondents found it difficult to make a choice and reasons for choosing one version over another were based on marginal differences between two forms. A few respondents could not make a choice between the two formats that they completed.

It is also interesting to compare how specific formats of the census form compared with one another. The following results summarize the highlights of these comparisons:

- Percentage of times that the ***sequential format*** was preferred to the ***matrix format* = 65%**
- Percentage of times that the ***book format*** was preferred to the ***single sheet format* = 58%**
- Percentage of times that the ***sequential book*** was preferred to the ***matrix book*** = 77%
- Percentage of times that the ***sequential sheet*** was preferred to the ***matrix sheet*** = 71%

Therefore, the findings clearly indicated that respondents preferred the sequential format over the matrix format and the book format over the single sheet format.

The think-aloud interviews determined many reasons for respondents' preferences. Their choices were usually based on very clear and well-reasoned decisions that took into account very practical considerations. For example, many preferred the sequential format because they felt that it was easier to read and complete than the matrix format. They noted that the questions followed a logical order with a separate section for each person and that the answer categories appeared with each question.

Respondents preferred the book format over the single sheet format because it was easier to handle. They noted that the single sheet format was "inconvenient" (*i.e.*, cumbersome and difficult to handle when open on a table) and that the book format looked "more official and serious." Many also liked the book format because the English and French questions were together on the same form as opposed to receiving two separate forms for each official language in the case of the single sheet format. Respondents who preferred the matrix book format often liked it because of the layout of its second two pages and because it was like the forms used in previous censuses.

Tables 4, 5, 6, and 7 summarize the strengths and weaknesses of the sequential, matrix, single sheet, and book formats as indicated by respondents during the testing of the four formats.

5. CONCLUSION

Although there were strengths and weaknesses noted in each of the four formats of the census form that were tested, the research clearly determined that respondents preferred the sequential format over the matrix format and the book format over the single sheet format. Among the combinations of sequential, matrix, single sheet and book formats that were tested, respondents indicated a preference for the sequential book format. Therefore, it was recommended that the sequential book format be used in the May 1997 Census Test with certain modifications. The recommended modifications took into account the strengths of the matrix format such as respondents' preferences for the layout on the second two pages of the matrix book. Even though the sequential book format was the preferred option of most respondents, the testing demonstrated that respondents were able to complete each of the four formats (matrix book, matrix sheet, sequential book, and sequential sheet) with relatively few problems. For each of the four formats tested, respondents were able to provide accurate responses. Therefore, this study has determined that any of the formats tested is a viable option for the census form, with each format having its strengths and weaknesses. The differences in respondents' preferences were based on issues such as the appearance of the census form, its readability, and ease of handling. In terms of respondent-friendliness, the sequential book format performed best.

6. ACKNOWLEDGMENTS

The author wishes to acknowledge the contributions of Dennis Cooley, Danielle Laroche and Jean-Pierre Morin of Statistics Canada who, along with the author, conducted the cognitive interviews during this research study.

REFERENCE

Statistics Canada (1997). Cognitive Interviews: Design of the Questionnaires for the May 1997 Census Test. *Questionnaire Design Resource Centre*, Ottawa.

Table 4
Sequential Format

Strengths	Weaknesses
The sequential format is easy to read. The questions follow a logical order (they "allow you to walk through the questionnaire"). The questions ask about one person at a time. The answer categories appear with each question.	The same questions, instructions and responses must be repeatedly read for each person. The sequential format seems longer than the matrix format. There is provision for only 5 persons on the sequential format (*vs.* 6 persons on the matrix format).

Table 5
Matrix format

Strengths	**Weaknesses**
The matrix format is the format that has been used in previous census forms (respondents, therefore, are familiar with it). The questions for all persons are on the same page. Some respondents prefer the matrix format because the answer categories are beside the questions instead of below them.	Some respondents find the matrix format confusing. The matrix is too large for some respondents to follow easily. The matrix format is not as readable as the sequential format. The matrix format requires more effort to understand and complete for some respondents. The questions appear to the left of and are separated from the answer categories. Left-handed respondents' hands cover the questions as they check or write their answers. Some respondents (especially older respondents) find the pre-printed response for Person 1 in Question 2 difficult to understand.

Table 6
Single Sheet Format

Strengths	Weaknesses
Some respondents prefer the census form on only one sheet. The single sheet format is compact in size. The single sheet format is shorter than the book format. There are fewer pages to turn than the book format. The single sheet format is visually better than the book format because everything is on the same page. Respondents can see everything at once. They can see the end of the questionnaire when beginning Step 2. The list of household members in Step 2 can be seen when completing Question 1. Separate forms and different colours for English and French are liked by many respondents.	Many respondents prefer the census questionnaire in a book format ("I expect a book...people are a little more accustomed to a book...I tried to turn the pages like a book"). The single sheet format is less official/serious than the book format (it "looks like a brochure"). The single sheet format is cumbersome and difficult to handle when open on a table ("It's bothersome...the length of the page, flipping it open halfway across the table and having to fold it"). The single sheet format is "inconvenient" ("To find a space where I could open this up and actually work would be a miracle in my household some days....I just don't like fold-out things!"). Folding is annoying to some respondents ("It's like a road map"). Separate forms for English and French are "a waste of taxpayer's money."

Table 7
Book Format

Strengths	Weaknesses
The book format looks more official and serious than the single sheet format. The book format is easier to handle than the single sheet format. The book format is "a lot neater" than the single sheet format. The form can be folded and placed in the return envelope more easily than the single sheet format. English and French are on the same form. The bilingual format is more economical and practical. Many respondents like different colours for English and French.	The book format is bulkier than the single sheet format. The book format looks longer than the single sheet format. The book format uses more paper than the single sheet format.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

QUALITATIVE TESTS AS A SUBSTITUTE FOR EXPERIMENTAL PILOT STUDIES TO COMPARE QUESTIONNAIRE VERSIONS

A. Luiten, I. Kuijpers and H. Akkerboom[1]

ABSTRACT

This paper focuses on the combination of "qualitative testing" with "down-scaled split-ballots", making proper use of computerized testing. The first case study shows how a small-scale test by 'in-depth interviews' helped design a medium-scale test, in which 'meta-interviews' were held with representatives of an ordinary sample of government institutions, who were asked to report on Vacancies Difficult to Fulfill (*VDF study*). The result was that the quantitative study originally planned was cancelled. The second case study was a medium-scale 'in-depth comparative test' of Labour Force Survey (LFS) questionnaire versions in three neighbouring European countries (with two languages in one country). The 'national questionnaires' were contrasted with an 'ideal' version acting as a control in paired comparisons (*LFS-Euregio study*). Here the main trick was to use proxy information to produce a second series of data on the same test subject. The result was a calibration questionnaire which can be used in a quantitative follow-up study. The third case study, still in progress, also concerns the LFS: certain LFS questionnaire design principles have been tested in a two-phase experimental setup, with small-scale qualitative comparisons being used to design medium-scale ones (*LFS-Eurostat study*).

KEY WORDS: Questionnaire; Qualitative tests; Experiments; Meta-information; Respondent consultation.

1. INTRODUCTION

In questionnaire design and development, a preferred strategy of asking questions, collecting answers, and processing data will guide the steps taken to design and implement the survey in question. Nevertheless one will think over various alternative formulations of concepts and questions, and various possibilities for the order of survey topics, order of questions per topic, and answer formats. Similarly, design decisions concern the various modes available to recruit sample respondents, administer the questionnaire, and so on. Many choices will be made informally, but for some issues there may be need and opportunity for *Questionnaire Testing (QT)*, say in small-scale 'laboratory studies' involving $\underline{v}$ = 5 - 50 test subjects ($\underline{v}$ = volunteer respondent) and/or $\underline{i}$ = 5 - 30 interviewers. Similarly, consultation of all kinds of experts may provide evidence to improve on the choices to be made.

A few major design issues may require thorough research in some *experimental pilot study*, designed to yield statistically significant conclusions as to 'which of two or more design versions produces the best answers, the best response rate, *etc.*' Usually such an experiment has the character of a 'quantitative comparative field test', of which 'split-ballots' are a well-known example. Such a study requires properly designed samples of moderately large size (say $\underline{n}$ = 150 - 1500) and tight control of the fieldwork. This makes experimental pilot studies generally expensive and difficult to perform.

In this paper we discuss three case studies in which an experimental pilot study was the ultimate aim, but some sort of *preparatory qualitative testing program* was carried out first (mainly in the field); in all cases, so-called "Qualitative Content Tests" formed the core of the program, mostly preceded by some "Definition Study" meant to research the feasibility of the tests and prepare for them, *cf.* Akkerboom and Dehue (1997, Table 1a).

2. DEVELOPMENT OF THE SURVEY INSTRUMENT

2.0 Definitions

The commonest method in so-called "Qualitative Content Tests" is the *one-to-one meta-interview* (sometimes called a 'cognitive test'), which focuses on the questionnaire (including its information) and the test person's understanding of it, their attitudes towards survey topics and concepts, their memory processes and ability to respond, *etc.* In Akkerboom and Dehue (1997, Table 1b), a '*meta-interview*' is defined as an *ordinary 1-1 interview supplemented by meta-interviewing techniques*. Examples are (re-)interviews with techniques like (a) focussed probes and/or (b) vignette tasks and/or (c) sorting tasks and/or (d) paraphrasing, possibly combined with (e) behavioural observation. *Focussed probes* can be (a1) comprehension probes, (a2) reference set probes, (a3) explanatory or expanding probes (answers), (a4) check questions, (a5) confidence or sensitivity ratings, *etc.* Focussed probes often go together with (f) *general prompts* to elicit respondent comments or remarks relevant to what (s)he thinks, does

[1] Hans Akkerboom, Questionnaire Design Resource Centre, Department of Data Collection Methodology (GWM), Statistics Netherlands (StatNeth), P.O. Box 4481, 6401 CZ Heerlen, The Netherlands; e-mail: HAKM@CBS.NL, hans.akkerboom@gironet.nl.

The views and opinions expressed in this paper are those of the authors and do not necessarily reflect official Statistics Netherlands (NCBS; here we use the abbreviation 'StatNeth') policies.

and feels while answering the survey questions. Other methods often used in Qualitative Content Tests are the (respondent) *focus group*, which is essentially a 'many-to-one open interview', and the *in-depth interview*, which is the analogous 'one-to-one open interview'. In Akkerboom and Dehue (1997, Table 1b), a *'focus group'* is defined as a *many-to-one (n-1) open meta-interview with a topic list as agenda*. Focus groups, or in-depth interviews, are relatively unstructured and exploit the creative reactions of the (group of) respondents. The topics of the meta-interview are determined by the questionnaire prototype and, partly, by suspected response error risks.

Thus Qualitative Content Tests yield *respondent-related meta-information*, that is empirical information on various quality aspects of the questionnaire, in addition to the information that the survey is supposed to collect. Correspondingly, a *meta-question* (focussed probes, general prompts, *etc.*) is a question about any kind of problem that test subjects may have with one or more 'regular questions'.

2.1 VDF study: "Defining and Measuring Vacancies Difficult to Fulfill"

2.1.0 VDF Study: Overview

The *first case study* (*VDF: costs of about $46,000*) shows how a small-scale test by 'in-depth interviews' helped design a medium-scale test, in which 'meta-interviews' were held with representatives of an ordinary sample of government institutions, who were asked to report on Vacancies Difficult to Fulfill (*VDF study*). The study was set up to explore subjective versus objective ways to ask for VDFs. The original plan was to carry out a quantitative follow-up study after the small-scale and medium-scale qualitative tests. This study would have measured a new 'objective VDF item battery' along with the old 'subjective VDF question'. This should have yielded an estimation procedure by which corrections could be computed for a series of outcomes of past surveys, featuring only the subjective question. However, the follow-up study was cancelled due to the findings of the two qualitative tests.

2.1.1 VDF: Review of the Client's Problem

Every quarter, Statistics Netherlands sends a questionnaire form to government organizations to measure their number and nature of vacancies. Respondents usually belong to the personnel department, where they are responsible for administration of relevant data like date of hiring, number of hours worked per week, type of contract, etc. The questionnaire form gives an explicit definition of what a 'vacancy' is supposed to be. Once a year a few additional questions are asked about the nature of these vacancies. One of these is a simple subjective question, by which respondents are asked to check for every (type of) vacancy listed whether "You consider this vacancy difficult to fulfill." StatNeth was sponsored to

- *Evaluate the subjective VDF question:* evaluate whether the present subjective question about Vacancies Difficult to Fulfill yields reliable and consistent interpretations and perceptions, while no uniform definition of VDFs is presented to the respondent (for various 'sectors' of government);
- *Design and test a new 'objective VDF item battery'*, meant to be asked in addition to the subjective question, so that labour market conditions, in particular difficulty of hiring personnel, can be validly and consistently compared between government sectors.

The ultimate aim was to carry out a quantitative follow-up study in which the additional objective VDF items had to yield an estimation procedure, by which corrections could be computed for a series of outcomes of past surveys, featuring only the subjective question.

A review of the client's problem yielded a number of factors that could influence reliability and consistency of the subjective VDF question:

1. The *interpretation of 'vacancy'* by itself;
2. The *administrative procedures* influencing the listing and rating of vacancies;
3. The *potential interaction between VDF perception and VDF causes and solutions*.

2.1.2 VDF: The Questionnaire Testing Program

Apart from a 'review' by a number of experts, qualitative tests were obviously necessary to develop a good instrument to measure VDFs objectively and to explore any risk factors for the ultimate quantitative follow-up study. Between March 1996 and February 1997 StatNeth performed the testing program described in Table 1.

Initially, the purpose of Step 3 had been to test candidate VDF items by themselves, using rather focussed meta-questions, and not by a comprehensive vignette study. However, Step 2 did not yield unequivocal candidates for objective VDF items, because

- VDF perception turned out to depend heavily on interpretation of the 'vacancy' concept,
- respondents do not have all the necessary information at their disposal and so make 'best guesses',
- VDF perception was heavily confounded with perceived causes and solutions.

Some details of the checklists typical for the in-depth interviews in Steps 1 and 2, and the meta-interviewing techniques typical for Step 3, are given in Table 2.

2.1.3 The VDF Study Outcome

As a result of Step 3, three items were considered good candidates for objective VDF measurements (the number of 'adequate' candidates reacting on a recruitment action, the need to have a second round of recruitment, and a measure of vacancy duration, if that could be suitably standardized for between-sector differences). However, the administrative problems that had been discovered ("which person to contact best, and how?"), and the inconsistent interpretation of the 'vacancy' concept by itself, had put serious doubts on the feasibility of the original plan to carry out a quantitative follow-up study.

Table 1
The VDF Questionnaire Testing Program ($\underline{v}$ = volunteer respondent, $\underline{n}$ = sample respondent).

Step	**Topics** per step	**Methods** per step (**test size**)	**Main Test Tools**
Step 1 (Definition and feasibility study)	Evaluation and exploration of *key concepts* ('vacancy' and 'filling difficult vacancies'). Exploration of *administrative procedures.* Preliminary inventory of causes and solutions of VDFs.	**In-depth interviews** (**$\underline{v}$ = 15, 'at location'**), with a few meta-questions. *Institutions were purposively selected* from the most important 'sectors' of government ('over sampling' of institutions with VDFs). *Employees* were interviewed who had filled in the form in the past year as a survey participant.	- *Topic list* according to review (v.2.1.1). - *Open discussion* triggered by completed form. - *Spontaneous comments* w.r.t. definition, comprehension, *etc.* - *Paraphrasing.*
Step 2 (Qualitative Content Test)	Evaluation of *candidates for objective VDF measures.* Evaluation of *operational needs and possibilities.* Construction of the *meta-questionnaire* for Step 3.	**Exploratory focus group** (**$\underline{v}$ = 1 × 4**) with personnel advisers who were no respondents to the regular survey. **In-depth interviews** (**$\underline{v}$ = 5 at location**) with employees like in Step 1, with new form/item battery.	- *Open discussion* on VDF measures, administrative procedures, roles of respondents. - *Focussed probes.*
Step 3 (Qualitative Content Test)	Inventory of *interpretations* of vacancies, *perceptions* of VDFs. Test of *candidate VDF items.* Final inventory of causes and solutions of VDFs.	**Meta-interviews** (**$\underline{n}$ = 150; divided equally over CAPI and CATI, at location**) with a (sector, size) stratified sample of employees/institutions. **Evaluative focus group** (**$\underline{v}$ = 1 × 6**) with employees.	- *Vignettes* on vacancies and VDFs. - *Focussed probes.* - *Confidence ratings of answers.*

2.2 LFS-Euregio study: "Worst Case Comparison of Labour Force Survey Questionnaire Versions"

2.2.0 LFS-Euregio Study: Overview

The *second case study* (*LFS-Euregio: costs of about $27,000*) comprised a medium-scale 'in-depth comparative test' of Labour Force Survey (LFS) questionnaire versions in three neighbouring European countries: Belgium, Germany, and The Netherlands (versions in two languages in Belgium). The 'national questionnaires' were contrasted with an 'ideal' kind of 'calibration questionnaire', acting as a control in paired comparisons. We could rely on qualitative methods, in particular 'reconciliatory debriefings', to ensure the plausibility of this method. The comparisons proved a further experimental pilot study to be feasible and necessary to produce 'correction factors' to account for differences in national LFS outcomes. Moreover, the qualitative study showed the issues to be addressed in such an experiment.

One could have considered to omit the Qualitative Content Test and start with a proper quantitative comparison by itself. However, this would not have obviated the need for design and development of a calibration questionnaire. Apart from that, without a Qualitative Test, the experimental design would have to include at least eight independent sub-samples to allow separate comparison of the calibration questionnaire with the questionnaires in each of the four languages (n = 800 was the total size originally planned for inclusion of four urban areas in the three countries: a very large study to begin with).

2.2.1 LFS-Euregio: Review of the Client's Problem

The EMR, or "Euregio Meuse Rhine", is a region where three European countries meet around the cities of Aachen, Hasselt, Liège, and Maastricht. The EMR is also an institution for consultations between people who cooperate over the border. To many people, an important issue is to know whether figures on inactivity, unemployment, and employment (according to the definitions of the International Labour Organization) can be interpreted in a way that a consistent overall picture for the EMR arises. StatNeth was sponsored to investigate the *feasibility of producing correction factors* for the national estimates, so as to ensure comparability of (future) regional statistics. The available budget was just sufficient for a preparatory qualitative study, which had to answer the following questions:

- What factors, in terms of population characteristics and questionnaire characteristics, are crucial to comparability of Euregional labour status statistics?
- Would it be feasible to harmonize national questionnaires by some 'calibration questionnaire'?

2.2.2 LFS-Euregio: The Questionnaire Testing Program

The Euregio study, carried out between January and December 1996, included test subjects from the four different language/culture groups. A Questionnaire Testing Program was worked out in three steps, see Table 3. For 24 persons to be classified according to ILO definitions, the performance of the national questionnaire was compared to an 'ideal' calibration questionnaire especially constructed for the purpose. Most interviews were double interviews, both with the respondent to be classified and with one of his or her intimate proxies (partners, parents). One of the two test subjects was administered the national questionnaire, the other one the calibration questionnaire: labour status was assessed for one of the two only, the same person in both interviews. At the end of the two separate interviews, both respondents were involved in an in-depth interview, or 'reconciliatory debriefing', that focussed on differences in the two ILO classifications for the same person, and on possible explanations of differences in the interpretation of key items in one or the other questionnaire. The reconciliatory debriefing was also used to investigate whether differences were artificial in the sense that they were due to differing informants (proxy partner or the subject of the questions him/herself). Hardly ever did this procedure lead to serious 'marital fall out' or the like. In a few cases, when an intimate proxy was not available, two questionnaires were administered to the same person successively, followed by a similar kind of reconciliatory debriefing. Both procedures could be checked to be feasible and reliable: there was no evidence of serious order effects that could invalidate conclusions as to differences in the questionnaire versions compared. All interviews were computer-assisted, with ILO derivations and focussed probes being programmed along with the two question-naires, while the in-depth techniques of the reconciliatory debriefing were guided by a topic list only.

Table 2
VDF examples of checklists, meta-interviewing tools, and meta-information obtained

Steps	Examples of checklists	Meta-interviewing tools	Meta-information obtained
Step 1	1. Do you use other information sources while filling in the form? *If "Yes",* - which sources? - for what information? - or which informants?	• explanatory and expanding probes, general prompts, *etc.*	1. Usually employees do not make use of other information sources or informants. They use a datafile in which personnel changes are registered. The file gives an overview of where and when vacancies arise, when they are fulfilled. Information on substantial issues about those vacancies can only be obtained by contacting personnel advisers.
	2. What does 'the end of the quarter' mean to you in this question? For which months did you fill in the form?	• comprehension and reference set probes, *etc.*	2. The last day of the month, the thirtieth day. But the reference date for calculations is the third Wednesday of the month.
	3. Can you formulate the following question in your own words: "For how many vacancies do you search for personnel during less than 3 months'?	• paraphrasing	3. Here duration is interpreted in three different ways, namely a) How long one is (already) busy searching personnel. b) How long one expects to be busy searching personnel. c) How long one wants to hire the personnel being searched.
Step 2	1. Is the form received by the right person in your organization? (The right person is the person who can answer the questions on the form.) - Do the employees have the information asked or do they consult another person? And for what questions? - Which information is not available?	• all kinds of focussed probes and general prompts	1. For relatively small organizations it makes no difference which person receives the form. What matters is that the form goes to the personnel department; necessary information is always easily available. For relatively large organizations an employee of the personnel department receives the form. This person has information only about the number of vacancies, not about further details. For these organizations is it almost impossible to acquire specific information about the employees from all the personnel advisers in various (sub-) departments.
	2. How many vacancies require a second recruitment round? - Is this for you a criterion to distinguish a VDF? - What is a "second recruitment round"?	• item battery test: check questions, comprehension probes, *etc.*	2. It is easy to deliver this information, because vacancies are published by the personnel department. When in the first recruitment round everything is done to fulfill the vacancy, then a vacancy for which a second recruitment round is needed is recognized as a VDF. Second recruitment means that a second (internal or external) round is started.
Step 3	1. Did you (would you) count an entry on the present form as a vacancy if it concerns a case released for external recruitment?	• vignette tasks	Counts as a vacancy 98% (n=150) Doesn't count as a vacancy 2% (n=150)
	2. How certain are you about your answer?	• confidence ratings	*Value Label* 20-40% 40-60% 60-80% 80-100% Missing *Frequency* 1.4% 9.6% 17.8% 69.9% 1.3% *(Total n= 73)*

Table 3
The LFS-Euregio Questionnaire Testing Program (e = expert, v = volunteer respondent).

Step	Topics per step	Method per step (test size)	Main Test Tools
Step 1 **(Definition and feasibility study for Steps 2 and 3)**	Design of *calibration questionnaire* and *research strategy* (paired comparisons; double interviews with reconciliatory debriefings).	***Review and expert appraisal (e = 5)***	*Desk research* into differences between national survey methods (questionnaires, fieldwork methods, sampling, *etc.*)
Step 2 **(Qualitative Content Test)**	Interpretation of items and answers. *Comparability of translated calibration questionnaires*. Operationalization of research strategy. Construction of Step 3 *calibration questionnaires.*	***Meta-interviews (v = 5)*** with test subjects from 'worst case subpopulations'. ***Expert re-appraisal (e = 2).***	- *Focussed probes.* - *On-line derivation of ILO labour status (of the one person that was the subject of the questions), as seen by (both) person(s) interviewed.* - *Open reconciliatory debriefing.*
Step 3 **(Qualitative Content Test)**		***Meta-interviews*** ***(v = 20 + 20 in paired comparisons, followed by reconciliatory interview;*** ***v = 4 sequential comparisons plus reconciliation;*** ***v = 8 one-to-one)***; dealing with 'worst case subpopulations'	- *Focussed probes.* - *On-line derivation of ILO labour status, as seen by (both) person(s) interviewed.* - *Open reconciliatory debriefing.*

Respondents were recruited from subpopulations expected to be at risk of erroneous measurement. Persons belonging to the core of the working population were ignored. The focus was on recruitment of persons with irregular or incidental jobs, jobs without a contract, minor jobs, jobs from which they were temporarily absent, people with unpaid work (*e.g.*, volunteers), people helping in a family business, *etc*.

Some details of the reconciliation techniques and focussed probes used in Steps 2 and 3 are given in Table 4, which also gives an idea of the kind of meta-information obtained.

2.2.3 The LFS-Euregio Study Outcome

The main study outcome was that the calibration questionnaire correctly identified people's 'true labour status': people in 'worst case situations' were correctly classified. The ILO classifications obtained by the national questionnaires were incorrect in a few cases, the important point being that they deviated from the calibration results in varying ways. In particular, different categories of people were measured sub-optimally in different countries. Other differences between countries concerned (1) the percentage of proxy answers in the national LFS and (2) the number of misinterpretations of individual key items in the national questionnaire. The second problem did not affect the ILO classifications very much, but more detailed data on the kinds of jobs that people have or that they search for would be seriously affected. In general, the qualitative LFS-Euregio tests proved the feasibility and desirability of an experimental study for calibration of national figures, for which funds still have to be raised.

2.3 Eurostat-LFS Study: "Probing into Rules for LFS Questionnaire Structure"

The *third case study* (*LFS-Eurostat: costs of about $110,000*) concerns a two-phase experimental setup, with small-scale comparisons being used to design medium-scale ones. Here ca. 50 meta-interviews in the first steps (a Qualitative Content Test in the StatNeth questionnaire laboratory) have led to the formulation of sensible hypotheses, which thereafter have been researched qualitatively in the field through ca. 200 meta-interviews, carried out in the same countries as in the LFS-Euregio study (a kind of Qualitative Experiment: qualitative experimental measures serve to compare a controlled administration of various questionnaire versions, which were carefully designed in the first steps). Global issues of structure, which involve several questions at a time (order and context of questions and question groups), were mainly researched by between-subjects designs, whereas 'local issues of formulation', which mainly involve separate questions wording, format, introduction, *etc.*) were treated by within-subject check questions and the like. This study, still in progress, is expected to produce empirical evidence that can help members of the European Union to agree on a set of 'target structure principles'. According to such principles, questionnaire design for the 'European Labour Force Survey' can be harmonized.

Table 4

LFS-Euregio examples of meta-interviewing tools and meta-information obtained.

A. Focussed probes and spontaneous respondent's remarks

A lot of problems with concepts or questions were unearthed by spontaneous respondent remarks. The answers showed that several key concepts were interpreted in a way that could potentially influence ILO classification. To illustrate, here is part of an interview with a farmer, working 97 hours a week on his farm and in an additional cattle haulage firm.

INT: Do you consider yourself primarily as working with paid work?
RESP: It's my company...but paid work.... what I earn stays in the company, so I'd say 'no'.
INT: Do you have any paid work at the moment, even for one hour per week or a short period of time?
RESP: No... what do you mean, paid work. I only have the work I do now, no more.

Had the respondent just sticked to yes/no answers, without clarification, he would have been classified as ***'unemployed'***.

B. Reconciliation interview

Usually, the research subject (RS) and his /her proxy (PR) had different opinions on one or more questions. The last part of the interview was meant to find out who was 'right' (not at all always RS!). For example, here is part of an interview where RS said she works and has an employment contract, whereas her husband maintains that she doesn't. The interview shows that the husband is right and that RS is actually unemployed.

INT: Things began to diverge the moment you said your wife did not have a contract.
PR: No, she stopped working for a year and she had to give up her contract, in order to be able to do it.
RS: I just feel I am on leave
PR: No, you had to sign a letter of resignation
RS: But they guarantee that I can come back
INT: Does it feel like you still have a job?
RS: Yes, I feel like, I have a job and, at the same time, I am on leave
INT:But officially, you don't have a job?
PR: No, officially not.
RS: But they sent me a letter that I can start work again by February.
PR: You have to be honest, had she asked you to show her the contract, you could not have produced it.
RS: Well OK, if you really want to be this precise...

Ultimately, however, the ILO classification of this RS was the same ('working') in both interviews, because proxy, in the calibration survey, mentioned hours his wife worked as a piano player. This is an example of a case where quantitative agreement in classification can mask fundamental differences in background of that classification.

Sometimes, if the interviewer suspected that partners would give different answers, she would ask questions that one of the partners had not encountered in the previous interview, for example:

INT: The first question I asked your wife was: Does your husband consider himself as a working person with paid work. What would you have said?.

At times, respondents could be very surprised by answers given by their partners:

INT: Would you like to have work for 12 hours or more per week?
RS: oh, yes!
INT to PR: you said that she would not, didn't you?
RS: I surely would like that.
PR: Why would you like to do that, to earn money, or as a way to spend your time?
RS: Both, I would like to work with aged people, but I don't want to do it as a volunteer. I want to be paid for my efforts.
INT: Your husband seems to be flabbergasted!
PR: Would you really want to work for money?
RS: Yes.
PR: You're mad!
RS: Yes, I know, but still.
PR: Well, you know how I feel about it.

3. DISCUSSION

In all case studies, the first step of the testing program took place under 'laboratory conditions', whereas the following step(s) used qualitative research methods in experimental field interviews. In all laboratory tests, recruitment of test subjects was based on availability and suitability for the specific test goals. In the qualitative field tests, a systematic sample was drawn in one case, whereas the other two cases used a quota scheme to recruit test subjects from very special subpopulations. In all cases, the collection and analysis of experimental data is qualitative: the main yardstick of comparison is the 'meta-information' about 'data quality' rather than the data themselves. The first steps of the studies had been conducted by questionnaire testing experts. Computerized testing was of great help in these phases, because the various 'experimental conditions' could be properly implemented. The meta-interviews in the last step of both the VDF study and the LFS-Eurostat study were conducted by specially trained interviewers whose usual job is to do regular survey interviews in the field. The most obvious disadvantage of the case studies was, of course, that they did not permit any reliable conclusions as to the impact of their findings on proper survey outcomes. However, comparison of questionnaire versions by (preparatory) qualitative tests generally appears to have the following advantages:

- relatively low costs, as compared to experimental pilot studies,
- easy control of experimental conditions, as far as questionnaire versions are concerned,
- dynamic design options: outcomes of one step determining the setup of the next one (*e.g.*, timely rethinking of the need for a quantitative follow-up study), *cf.* Akkerboom and Luiten (1997).

ACKNOWLEDGMENT

The first two authors work as qualitative researchers in the StatNeth Questionnaire Design Resource Centre, the third one as the QDRCs program manager. All three shared most of the responsibilities for the case studies LFS–Eurostat, VDF, and LFS–Euregio discussed here. They want to acknowledge Math Bosch, Gé Conen, Francine Dehue, Hans van Kerkoerle, Frans Kerssemakers, and Pieter Stel for their contributions.

REFERENCES

Akkerboom, H., and Dehue F. (1997). The Dutch model of data collection development for official surveys. *The International Journal of Public Opinion Research*, 9 (2), 126-145.

Akkerboom, H., and Luiten A. (1997). Selecting pretesting tools according to a model of questionnaire development, with illustrations concerning patient satisfaction with medical care. Paper presented at the 1996 AAPOR Conference, May 16-19, Salt Lake City, Utah, United States; ASA/AAPOR *Proceedings Survey Methods Section*, 911-916.

INTERPOLATING TIME SERIES FROM ADMINISTRATIVE DATA WITH DIVERSE REFERENCE PERIODS

P.A. Cholette[1] and J.-L. Coster

ABSTRACT

Data supplied to statistical agencies often cover fiscal years, fiscal quarters, fiscal months (a number of weeks), instead of calendar periods. The data in the Canadian Goods and Services Tax administrative file cover, to all intents and purposes, any number of days. This paper presents a method to calendarize such data. The method first produces daily interpolations which comply to the fiscal data. Calendarized values are then obtained by summing the interpolations over the desired calendar periods. By taking advantage of the very diversity of reporting periods, the method can produce monthly estimates, based only on fiscal quarter data (say), without the benefit of monthly data.

KEY WORDS: Benchmarking; Interpolation; Irregularly spaced data; Recursive least squares.

1. INTRODUCTION

Very often data cover fiscal periods instead of calendar periods. Thus, the fiscal year of the Canadian federal and provincial governments ranges from April 1 of one year to March 31 of the following year; the fiscal year of many educational institutions, from September 1 to August 31; and that of banks, from November 1 to October 31. The fiscal quarters of these institutions correspond to their fiscal years, for instance November 1 to January 31, February 1 to April 30, and so forth. Retail and wholesale trade data often cover fiscal months, that is a number of consecutive weeks. For the sake of generality, this document assumes that the reference periods range from any date to any date, with no repetitive pattern, and thus cover any number of days. This assumption accommodates all possible situations, namely that of the administrative file of the Canadian Goods and Services Tax (GST), which is characterized by a multitude of fiscal periods.

This paper presents a calendarization method designed for situations such as that of the GST. Like in many such methods (see Cholette and Dagum 1994), the strategy consists of interpolating more frequent (*e.g.*, daily) values and then of taking the temporal sums over the desired calendar periods, whether months, quarters, *etc*. The method proposed is as a multivariate and non-linear generalization of the Chow and Lin (1971) method. The latter is widely applied to interpolate monthly series between yearly or quarterly data, using relevant monthly series as regressors and predictors. The model presented uses functions of time as regressors; it assumes that, within a chosen class of the population (*e.g.*, a given industry and a given province), the target interpolations of each individual business share a basic multiplicative seasonal pattern, but follow their own trend-cycle and disturbance components. The model is thus similar to panel data models (*e.g.*, Pfeffermann and Bleuer 1993), where a series is specified for each panel to have its time series components. The model can be viewed as a multivariate generalization of benchmarking methods with bias parameters (*e.g.*, Durbin and Quenneville 1997).

Section 2 presents the interpolation model. Section 3 outlines the solution based on linearized regression and recursive least squares. Section 4 presents a real case application. Section 5 discusses some issues related to the method.

2. THE INTERPOLATION MODEL

This section presents a multivariate model to interpolate time series from data with many reference periods. The *time resolution* $t = 1, 2, \ldots$ of the interpolations is monthly, if the month is the largest common denominator between of all the fiscal periods; and daily, if the day is the largest such denominator. Unless otherwise indicated, the interpolations are generated at the level of the individual businesses, $k = 1, \ldots, K$, providing the data.

The series to be interpolated contain the usual components of time series:

$$\theta_{k,t} = c_{k,t} \times \{s_t \times d_t\} \times e_{k,t}\,, \tag{2.1a}$$

or

$$\theta^*_{k,t} = c^*_{k,t} + \{s_t^* + d_t^*\} + e^*_{k,t},\, t = t_{F,k}, \ldots, t_{L,k};\quad k = 1, \ldots, K, \tag{2.1b}$$

where the asterisks denotes the logarithm. The trend component, $c_{k,t}$, the seasonal component s_t, the trading-day component d_t and the error $e_{t,k}$ are multiplicatively related. Within a given class (*e.g.*, one province, one

[1] Pierre A. Cholette, Statistics Canada, Time Series Research and Analysis Section, R.H. Coats Building, 3rd Floor "H", Ottawa, Ontario, Canada, K1A 0T6; e-mail: Cholette@statcan.ca. Jean-Louis Coster, Institut National de la Statistique et des Études Économiques, France, on sabbatical at Statistics Canada during the development of the method.

industry) all individuals share the same basic *seasonal-trading-day component* $\{s_t \times d_t\}$, whereas the other components are specific to each individual k. The time index $t = t_{F,k}, ..., t_{L,k}$ has no gaps but allows individuals to start and/or end their activity at different times.

The trend for each individual k is written as:

$$c_{k,t} = \alpha_{k,1} \times \alpha_{k,2}^{g_k(t)}$$
$$\text{or } c^{\bullet}_{k,t} = \alpha^{\bullet}_{k,1} + g_k(t)\ \alpha^{\bullet}_{k,2},\ t = t_{F,k}, ..., t_{L,k} \tag{2.2}$$

where $g_k(t)$ is a function of time such that the trend is a cubic constrained to flatten at two nodes located at the ends of the series to ensure a "conservative" trend.

The common seasonal and trading-day components are specified by means of H_s and H_d periodic functions. The seasonal is given by

$$s_t = \prod_{h=1}^{H_s} \beta_h^{p_h(t)}$$
$$\text{or } \ s_t^{\bullet} = \sum_{h=1}^{H_s} p_h(t)\,\beta^{\bullet}_h,\ 0 \le H_s \le 12 \tag{2.3}$$

where the functions, $p_h(t), h = 1, ..., H_s$, are a subset of the functions, $\sum_{\tau\in t} \cos(\lambda_i \tau)/m_t$, $\sum_{\tau\in t} \sin(\lambda_i \tau)/m_t$, $i = 1,...,6$, where τ is time in days, $\lambda_i = 2\pi i / 365.25$ and m_t is the number of days in period t. For the trading-day component d_t, the periodic functions, $p_h(t)$, $h = H_s + 1, ..., H$, $(H = H_s + H_d)$, are a subset of $\sum_{\tau\in t} \cos(\lambda_i \tau)/m_t$, $\sum_{\tau\in t} \sin(\lambda_i \tau)/m_t$, $i = 1, ..., 3$, where $\lambda_i = 2\pi i / 7$. (Details in Pierce, Grupe, and Cleveland 1984.) Note that all the β_h's are estimated at the daily resolution; this enables the calculation of the seasonal and of the trading-day components at the daily resolution, *i.e.*, s_τ and d_τ, regardless of the resolution of t. In the sequel, seasonality (or seasonal pattern) will refer to the aggregate $\{s_t \times d_t\}$, unless stated specifically.

The disturbance $e^{\bullet}_{k,t}$ of each individual is specified to follow a multiplicative seasonal ARMA model,

$$e^{\bullet}_{k,t} \sim \text{ARMA},\ E\left(e^{\bullet}_{k,t}\right) = 0,$$
$$E\left(e^{\bullet 2}_{k,t}\right) = \sigma_k^2,\ E\left(e^{\bullet}_{k,t}\, e^{\bullet}_{k',t}\right) = 0,\ k \ne k', \tag{2.4}$$

with parameters chosen by the statistician. Parameter σ_k^2 is estimated from the data.

The available data measure the appropriate temporal sums of the underlying interpolations:

$$y_{k,n} = \sum_{t\in n} \theta_{k,t} \equiv \left\{\sum_{t\in n} c_{k,t} \times s_t \times d_t \times e_{k,t}\right\},$$
$$n = 1, ..., N_k. \tag{2.5}$$

Note that (2.5) allows for *any* reference periods of the data, including gaps in the periods covered; and, that the interpolations are benchmarked to (*i.e.*, comply with) the observed value.

3. SOLUTION OF THE MODEL

This section outlines the solution for the interpolation model of section 2. Substituting the components (2.2) to (2.4) in (2.1b) and using matrix notation yields,

$$\theta^{\bullet}_{k_{T_k \times 1}} = X_{\alpha,k}\,\alpha^{\bullet}_k + X_{\beta,k}\,\beta^{\bullet} = X_k\,\gamma^{\bullet}_k + e^{\bullet}_k,$$
$$k = 1, ..., K,\ E\left(e^{\bullet}_k\right) = 0,\ E\left(e^{\bullet}_k\, e^{\bullet\prime}_k\right) = V_{e^{\bullet}_k}, \tag{3.1}$$

where

$$E\left(e^{\bullet}_k\, e^{\bullet\prime}_{k'}\right) = 0,\ k \ne k',\ \theta^{\bullet}_k = \left[\theta^{\bullet}_{k,t},\ t = t_{F,k} \cdots t_{L,k}\right],$$
$$\gamma^{\bullet}_k = \left[\alpha^{\bullet\prime}_k\ \beta^{\bullet\prime}\right]',\ \alpha^{\bullet}_k = \left[\alpha^{\bullet}_{k,1} \ldots \alpha^{\bullet}_{k,J_k}\right]'$$

where $X_k = [X_{\alpha,k} X_{\beta,k}]$ contains the explanatory variables of (2.2) and (2.3).

Equation (2.5) can be written as

$$y_{k_{N_k \times 1}} = L_k \exp\left(\theta^{\bullet}_k\right),\ k = 1, ..., K, \tag{3.2}$$

where L_k is a sum operator containing 0s and 1s.

The model consisting of (3.1) and (3.2) has a convenient recursive least square solution, which incorporates the data y_k of each individual separately:

$$\gamma^{\bullet(i)} = \gamma^{\bullet(i-1)} + V^{(i-1)}_{\gamma^{\bullet},\gamma^{\bullet}_k} Z'_{0_k}\left(Z_{0_k} V^{(i-1)}_{\gamma^{\bullet}_k} Z'_{0_k} + V_{0_k}\right)^{-1}$$
$$\left[y_{0_k} - Z_{0_k}\gamma_k^{\bullet(i-1)}\right]$$
$$V^{(i)}_{\gamma^{\bullet}} = V^{(i-1)}_{\gamma^{\bullet}} - V^{(i-1)}_{\gamma^{\bullet},\gamma^{\bullet}_k} Z'_{0_k}\left(Z_{0_k} V^{(i-1)}_{\gamma^{\bullet}_k} Z'_{0_k} + V_{0_k}\right)^{-1}$$
$$Z_{0_k} V^{(i-1)}_{\gamma^{\bullet}_k,\gamma^{\bullet}} \tag{3.3a}$$

$$\theta^{\bullet(\ell)}_k = \eta^{\bullet(\ell)}_k + V_{e^{\bullet}_k}^{(\ell-1)} L'_{0_k} V^{-1}_{0_k}\left[y_{0_k} - L_{0_k}\eta^{\bullet(\ell)}_k\right],$$
$$k = 1, ..., K;\ \ell = 1, 2, ...;\ i = (\ell - 1)K + k \tag{3.3b}$$

where

$$V_{0_k} = L_{0_k} V^{(\ell-1)}_{e^{\bullet}_k} L'_{0_k},$$
$$\eta^{\bullet(\ell)}_k = X_k\gamma^{\bullet(i)}_k,\ Z_{0_k} = L_{0_k} X_k,$$
$$y_{0_k} = y_k - L_k\theta^{(\ell-1)}_k + \quad L_{0_k}\theta^{\bullet(\ell-1)}_k,\ L_{0_k} = L_k \operatorname{diag}\left(\theta^{(\ell-1)}_k\right)$$

and where $\gamma^{\bullet(0)}$ and $V^{(0)}_{\gamma^{\bullet}}$ are starting values of the recursion. Vectors $\eta^{\bullet(.)}_k$ and $\theta^{\bullet(.)}_k$ respectively contain the *un-benchmarked interpolations* the *bechmarked interpolations*.

The recursion over individuals (k=1,...,K) is repeated for each *iteration* (ℓ ℓ=1,2,...) until convergence for all individuals. All parameters in $\gamma^{\bullet}$ change as a result of incorporating the data y_k of each individual in each iteration ℓ, this is why the superscript $(i) = (\ell - 1)K + k$ is used for $\gamma^{\bullet(i)}$, $V^{(i)}_{\gamma^{\bullet}}$ $V^{(i-1)}_{\gamma^{\bullet}}$. Note that $\gamma^{\bullet}_k = [\alpha^{\bullet\prime}_k\ \beta^{\bullet\prime}]'$, $V_{\gamma^{\bullet}_k}$

and $V_{\gamma^{*},\gamma_k^{*}}$ are not partitions but subsets of $\gamma^{*} = [\alpha_1^{*\prime} \ldots \alpha_K^{*\prime} \beta^{*\prime}]'$, $V_{\gamma^{*}}$ and $V_{\gamma^{*}}$ *sic* respectively. However some of the vectors and matrices pertaining to individual k change only once per iteration ℓ, this is why it is sufficient to use the superscript (ℓ), namely for the benchmarked and the *unbenchmarked interpolations* $\theta_k^{*(\ell)}$ and $\eta_k^{*(\ell)}$, respectively.

The "feasible generalized least square" algorithm to implement (3.3) consists of the following steps:

(1) Calculate:

$$L_{0_k} = L_k \operatorname{diag}\left(\theta_k^{(\ell-1)}\right), Z_{0_k} = L_{0_k} X_k, y_{0_k} =$$
$$y_k - L_k \theta_k^{(\ell-1)} + L_{0_k} \theta_k^{*(\ell-1)}.$$

(2) Calculate: $V_{e_k^{*}}^{(\ell-1)} = \sigma_k^{2(\ell-1)} \Omega_k$, $V_{0_k} = L_{0_k} V_{e_k^{*}}^{(\ell-1)} L_{0_k}'$, where Ω_k is the correlation matrix of the chosen ARMA process.

(3) Calculate:

$$\gamma^{*(i)} = \gamma^{*(i-1)} + V_{\gamma^{*}.\gamma_k^{*}}^{(i-1)} Z_{0_k}' \left(Z_{0_k} V_{\gamma_k^{*}}^{(i-1)} Z_{0_k}' + V_{0_k} \right)^{-1}$$
$$\left[y_{0_k} - Z_{0_k} \gamma_k^{*(i-1)} \right], \eta_k^{*(\ell)} = X_k \gamma_k^{*(i)} \theta_k^{*(\ell)} = \eta_k^{*(\ell)} +$$
$$V_{e_k^{*}}^{(\ell-1)} L_{0_k}' V_{0_k}^{-1} \left[y_{0_k} - L_{0_k} \eta_k^{*(\ell)} \right].$$

(4) Calculate: $e_k^{*(\ell)} = \theta_k^{*(\ell)} - \eta_k^{*(\ell)}$, $\sigma_k^{2(\ell)} = e_k^{*(\ell)\prime} e_k^{*(\ell)} / T_k$.

(5) Repeat steps (2) to (4) on the basis $\sigma_k^{2(\ell)}$ (instead of $\sigma_k^{2(\ell-1)}$

(6) Calculate:

$$V_{\gamma^{*}}^{(i)} = V_{\gamma^{*}}^{(i-1)} - V_{\gamma^{*}.\gamma_k^{*}}^{(i-1)} Z_{0_k}' \left(Z_{0_k} V_{\gamma_k^{*}}^{(i-1)} Z_{0_k}' + V_{0_k} \right)^{-1} Z_{0_k} V_{\gamma_k^{*}.\gamma^{*}}^{(i-1)}.$$

(7) After convergence, calculate:

$$V_{\hat{\theta}_k^{*}} = V_{e_k^{*}} - A_k V_{e_k^{*}} + B_k + A_k B_k A_k' - A_k B_k - B_k A_k'.$$

where $A_k = V_{e_k^{*}} L_{0_k}' V_{0_k}^{-1} L_{0_k}$, $B_k = X_k V_{\hat{\gamma}_k^{*}} X_k'$ where $\hat{\theta}_k^{*}$ and $\hat{\gamma}_k^{*}$ are the final estimates obtained at the last iteration. The square root of the diagonal elements of $V_{\hat{\theta}_k^{*}}$ are the coefficients of variations of the benchmarked interpolations.

Calendar month values (say) are then obtained by taking the monthly sums of the benchmarked daily interpolations: $\hat{c}_k = C_k \hat{\theta}_k$, $V_{\hat{c}_k} = C_k V_{\hat{\theta}_k} C_k'$, where C_k is a matrix of zeroes and ones, and the elements $[v_{t,r}]$ of $V_{\hat{\theta}}$ are given by $v_{t,r} = \{\exp(v_{t,r}^{*}) - 1\} \exp(\hat{\theta}_t^{*} + \hat{\theta}_r^{*}) = \{\exp(\hat{v}_{t,r}^{*}) - 1\} \hat{\theta}_t \hat{\theta}_r$.

4. EXAMPLE FROM THE CANADIAN VALUE ADDED TAX FILE

This section illustrates the interpolation methodology with a case from the Canadian Goods and Services Tax, namely to the sales by Hotel and Motels in Quebec, between November 1994 and February 1997. In this class, the file contained 21,617 records pertaining to 2,333 businesses. The interpolation was carried at a monthly resolution, after the businesses were aggregated into K=4 *fiscal groups*. These groups are the *monthly group* (referred to as "MM"), the *calendar quarter group* ("Q0"), and the two *fiscal quarter groups* ("Q1" and "Q2"), one ending in January, April, etc., and the other in February, May, etc. The recursive algorithm of section 3 was applied, with diffuse prior values for all the parameters and with a seasonal ARMA $(1,1)(1,1)_{12}$ disturbance. The regular and seasonal autoregressive parameters chosen are 0.90; and regular and seasonal moving average parameters, 0.50. The estimation took 7 iterations ($\ell = 1, \ldots, 7$); the convergence criterion required a maximum change lower than 0.1% in the interpolations of any group.

Figures 1 to 4 display the resulting monthly interpolations for each fiscal group, along with the data (divided by 3 and represented by steps for the quarterly groups). Each series of unbenchmarked interpolations $\{\hat{\eta}_{k,t}\}$ is the product of the estimated individual trend ($\hat{c}_{k,t}$) and the common seasonal pattern $\hat{s}_{t \times \hat{d}}$. Each series of benchmarked interpolations $\{\hat{\theta}_{k,t}\}$ complies in fiscal sum with the data ($\sum_{t \in n} \theta_{k,t} = y_{k,n}$). At the ends where there is no data, the benchmarked interpolations converge to the unbenchmarked ones, which can be viewed as the expected values of $\{\hat{\theta}_{k,t}\}$. The levelling-off of the trends at the ends ensure cautious extrapolations. The data of the quarterly groups display more seasonal amplitude than the common seasonal pattern (dominant in $\hat{\eta}_{k,t} = \hat{c}_{k,t} \times \hat{s}_t \times \hat{d}_t$). The departure from the common seasonal pattern, $\hat{s}_t$, is captured by the seasonal ARMA model: each $\{\hat{\theta}_{k,t}\}$ depart from its $\{\hat{\eta}_{k,t}\}$ in an repetitive manner from year to year as intended, including for the extrapolations. This desirable property is especially obvious in Figures 1 to 3 during the peak summer months; the reason would be that many quarterly reporters operate mainly in the summer. Thus the actual seasonal pattern has two parts: the part common to all fiscal groups and the residual ARMA part needed to accommodate each group.

Table 1 displays the estimates of the parameters (in the logs) with their approximate standard errors and t-ratios. The table also contains the standard-deviations of the ARMA residuals $\hat{e}_{k,t}$ (and not of the underlying noise) for each fiscal group. With smaller residuals, the monthly group (Sigma = 0.02395) had more weight in the calculation of the common seasonal pattern than the other more noisy groups.

Table 2 displays the seasonal pattern corresponding to the trigonometric parameters. The monthly seasonal pattern $\hat{s}_t$ consists of the geometric averages of the seasonal pattern $\hat{s}_\tau$ estimated at the daily resolution. We suspect the trading-day pattern $\hat{d}_\tau$ largely reflects the systematic error entailed in the formation of the fiscal groups; the grouping caused some loss of temporal resolution (discussed later).

Table 1
Parameter Estimates for Hotels and Motels in Quebec

Fiscal Group	Parameter Name	Value	Std. - Dev.	*t*-Ratio
MM	Cons	12.73395	0.00644	1978.67031
MM	Slope	-0.00477	0.00013	-35.74921
MM	Sigma	0.02395		
Q0	Cons	12.01379	0.033325	34.30338
Q0	Slope	0.00078	0.00071	1.10310
Q0	Sigma	0.11512		
Q1	Cons	11.2997	0.03469	325.76912
Q1	Slope	0.00199	0.00071	2.79330
Q1	Sigma	0.12112		
Q2	Cons	10.90546	0.01822	598.66706
Q2	Slope	0.00071	0.00037	1.91150
Q2	Sigma	0.06214		
all	CosS1	-0.16724	0.00221	-75.73848
all	SinS1	-0.08034	0.00222	-36.25323
all	CosS2	-0.00657	0.00174	-3.77521
all	SinS2	0.01965	0.00176	11.17335
all	CosS3	-0.00798	0.00180	-4.43488
all	SinS3	0.02175	0.00181	11.98316
all	SinS4	-0.02385	0.00192	-12.41558
all	CosS5	0.03254	0.00212	15.36759
all	SinS5	-0.01692	0.00217	-7.79889
all	CosS6	0.17850	0.01489	11.98405
all	CosD1	0.30090	0.01130	26.63733
all	SinD1	-0.22317	0.01196	-18.65471
all	CosD2	-0.40973	0.02624	-15.61501
all	SinD2	0.12535	0.02768	4.52794
all	CosD3	0.18074	0.03781	4.77964
all	SinD3	-0.21763	0.03053	-7.12768

Table 2
Monthly Seasonal Pattern and Daily Trading-Day Pattern

Month	Season. (%)	Cv (%)	Day	Daily (%)	Cv (%)
1	82.37	0.38	S	136.76	4.45
2	85.64	0.36	M	37.76	4.76
3	94.49	0.39	T	85.23	4.65
4	88.46	0.39	W	167.65	4.65
5	108.80	0.37	T	130.04	4.01
6	113.41	0.40	F	107.46	4.30
7	119.38	0.39	S	96.97	4.77
8	118.87	0.36			
9	116.12	0.39			
10	105.54	0.39			
11	90.01	0.38			
12	86.61	0.39			

5. DISCUSSION

This section discusses some of the problems raised by the interpolation model, namely the *statistical feasibility* of producing interpolations at high resolutions, and the *operational feasibility* of performing the calculations at the level of the individual businesses, especially at the daily resolution.

5.1 Resolution of the Interpolations

As mentioned, the method proposed can produce monthly interpolations in the absence properly monthly data; and daily interpolations, in the absence of daily data. One factor governing the resolution is of course the resolution of the prior information. For example occasional sample surveys could be designed to measure the seasonal and the trading-day patterns, which could then be used by the method in the form of prior information (The administrative data y_k would provide the level and trend-cycle of the target variable.) Another factor, which *can* be a substitute for prior information, is the variability of the reference periods of the data. Ideally, to interpolate daily, the data should cover a low number of days (not a multiple of 7) and be uniformly distributed throughout the days of the year. For the model such variability of the reference periods is a desirable attribute of the data – not a nuisance. A few clarifications in this regard are in order.

If all the data cover refer to fiscal years covering twelve consecutive months, none of the seasonal frequencies are observable, and matrices $X_k = \left[X_{\alpha.k} X_{\beta,k}\right]$ should not contain the regressors $X_{\beta,k}$. (Incorporating $X_{\beta,k}$ would cause singularities in (3.3).) Since there are no common parameters, the method can be applied separately for each individual, to produce non-seasonal monthly (say) interpolations. If the goal is to produce calendar year estimates (by taking the calendar year sums of the interpolations), these could be valid (or at least as reliable as possible with the method proposed), because seasonality cancels out on *any* consecutive twelve months. In other words, seasonality requires no modelling, because it is absent both from the data and from the calendarized values. If the goal is to produce quarterly interpolations, these would not be as reliable as calendar year estimates and they would be non-seasonal (*i.e.*, "seasonally adjusted").

Similarly, if all data cover fiscal and calendar quarters, the cosines and sines lasting three months (frequency $\lambda_4 = 2\pi 4 / 365.25$) and all the trading-day cosine and sines are virtually unobservable and should be omitted from $X_{\beta,k}$. The resulting monthly interpolations (say) would be seasonal but deficient at the omitted frequencies. If the goal is to produce calendar or fiscal quarter values, this deficiency poses no problem, because the missing frequencies (largely) cancel out on any consecutive three months. As in the annual case, the missing frequencies require no modelling, because they are absent from both the data and the calendarized values. For the same reason, yearly values would be even more valid. If the goal is to produce monthly interpolations, these could be mis-estimated, since there would be no trading-day component and the seasonal pattern is would be deficient at frequency λ_4. If this is not satisfactory, the estimated seasonal component could be removed from the interpolations to obtain non-seasonal monthly interpolations.

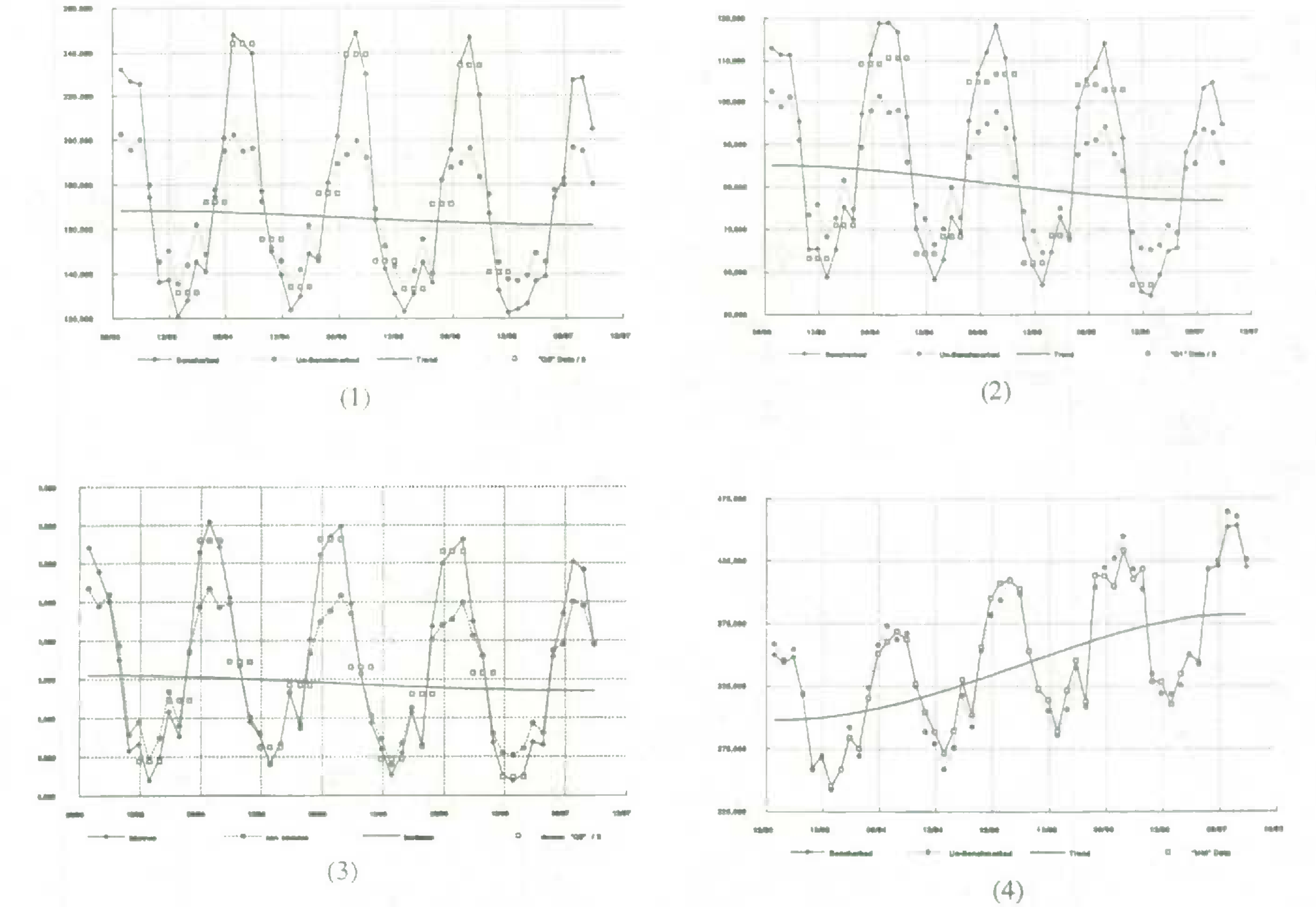
This discussion assumed that the data represented flows (*e.g.*, sales). If the data represent stocks (*e.g.*, inventories), referring to one month or one day of each fiscal period, they are highly seasonal (even in the fiscal year case) and require the seasonal regressors.

5.2 Operational Feasibility of the Calculations

As already mentioned, when the day is the largest common denominator of the various reference periods, the model should generally be applied with a daily resolution and most likely at the level of the individual businesses; the scale of the required calculations may then prove prohibitive. These two factors, time resolution and level of aggregation, govern the number of calculations. For instance if there are 2,000 businesses requiring each two individual trend parameters and 17 common seasonal parameters, the dimension of $V_{\gamma^\bullet}$ in (3.3) is 4,017×4,017 (129.09 Megabytes). If the method is applied to three years at a daily resolution, the dimension of matrices such as $V_{e_k^\bullet}$ and $V_{\hat{\theta}_k^\bullet}$ is 1,095×1,095 (9.59MB). The amount of memory required and the scale of calculations could then prove problematic, especially on a production schedule. To increase the feasibility of the method, simplifications to solution (3.3) are now outlined.

In order to reduce the level of aggregation, solution (3.3) is simplified by replacing $\gamma^{\bullet(.)}$ by $\gamma_k^{\bullet(.)}$, and $V_{\gamma^\bullet}^{(.)}$ and $V_{\gamma^\bullet}^{(.)}$ and $V_{\gamma^\bullet,\gamma_k}^{(.)}$ by $V_{\gamma_k}^{(.)}$; the largest dimension then goes from $(J+H)$ to (J_k+H) (*e.g.*, 4,017 to 19). This approximation should not seriously damage the estimates, because a large number of individuals implies small covariances between the parameters pertaining to different individuals.

In order to achieve a high resolution (*e.g.*, daily), solution (3.3) is simplified by using ordinary instead of generalized least squares. Matrices Ω_k and $V_{e^\bullet}^{(.)}$ respectively become I_{T_k} and $\sigma_k^{2(.)} I_{T_k}$. In fact there is no longer a need to store $V_{e^\bullet}^{(.)}$, L_k and L_{0_k}, because the elements of matrices such as V_{0_k}, γ_{0_k} and Z_{0_k} are known algebraically and can be calculated directly. The benchmarking operator $V_{e_k^\bullet}^{(.)} L_{0_k}' V_{0_k}^{-1}$ is known to have elements $1/m_n$, for t in the time periods covered by $y_{k,n}$, and 0 otherwise (where m_n is the number of periods covered). However this simplification would damage the quality of the interpolations, because it sacrifices the desirable properties of ARMA disturbances illustrated in section 4. One variant of this approach would be to apply the model only to obtain the high resolution seasonal parameters; the desired interpolations would then be obtained in a separate step. This second step would benchmark the seasonal pattern obtained in the first step,



(1)

(2)

(3)

(4)

Figures 1-4. Interpolations and Extrapolations for (1) the Calendar Quarter group, (2) the First Fiscal Quarter Group, (3) the Second Fiscal Quarter Group and (4) the Monthly Group using a univariate benchmarking method (*i.e.* Cholette and Dagum 1994, Durbin and Quenneville 1997), which is simpler but allows seasonal ARMA disturbances.

6. CONCLUSIONS

This paper presented a method to calendarize data with a wide variety of fiscal periods. The strategy consists of two steps.

a) The first step interpolates seasonal daily or monthly values between the fiscal data, under the assumption that within a class (*e.g.*, in a province and industry) all businesses share the same basic seasonal and trading-day patterns, and under benchmarking constraints imposing temporal additivity of the interpolations to the data.
b) The second step takes the temporal sums of the interpolated values over the desired calendar periods.

In many applications, the interpolations themselves are of primary interest. The interpolations may be monthly if the month is the largest common denominator between all the fiscal periods; and be daily, if the day is the largest such denominator.

In order to achieve interpolations with higher temporal resolution, the method depends on the very diversity of the fiscal periods, which turn out to be an essential attribute of the data and not a nuisance. As a matter of fact if the data does not display enough diversity of reference periods, some frequencies of the seasonal pattern are not observable; in such cases, section 5 points to the possibility of producing non-seasonal ("seasonally adjusted") interpolations.

Computationally, the method as presented has feasibility problems. This would occur if there are too many businesses in the class considered - especially at the daily resolution. In this regard, section 5 suggested approximations to the method currently being tested.

REFERENCES

Cholette, P.A., (1990). Transforming fiscal quarter data into calendar quarter values. (Eds) A.C. Singh, and P. Withridge, in *Analysis of Data in Time, Proceedings of the 1989 International Symposium*, Statistics Canada.

Cholette, P.A. (1990). L'annualisation des chiffres d'exercices financiers. *L'actualité économique*, 66, 219-320.

Cholette, P.A., and Chhab, N. (1991). Converting aggregates of weekly data into monthly values. *Applied Statistics*, 40, 3, 411-422.

Cholette, P.A., and Dagum, E. Bee (1994). Benchmarking time series with autocorrelated sampling errors. *International Statistical Review*, 62, 365-377.

Chow, G.C., and Lin, A.-L. (1971). Best linear unbiased interpolation, distribution and extrapolation of time series by related series. *Review of Economics and Statistics*, 53, 4, 372-375.

Chow, G.C., and Lin, A.-L. (1976). Best linear unbiased estimation of missing observations in an economic time series. *Journal of the American Statistical Association*, 71, 355, 719-721.

Di Fonzo, T. (1990). The estimation of *M* disaggregated time series when contemporaneous and temporal aggregates are known. *The Review of Economics and Statistics*, 72, 178-182.

Durbin, J., and Quenneville, B. (1997). Benchmarking by state space models. *The International Statistical Review*, 65, 1, 23-48.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 2, 163-177.

Pfeffermann, D., and Bleuer, S. (1993). Robust joint modelling of labour force series of small areas, *Survey Methodology*, 19, 2, 149-183.

Pierce, D.A., Grupe, M.R., and Cleveland, W.P. (1984). Seasonal adjustment of the weekly monetary aggregates: a model-based approach, *Journal of Business and Economic Statistics*, 2, 260-270.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A HISTORY OF EUROPEAN BUSINESS STATISTICS

F. de Geuser[1]

ABSTRACT

In general, statistical data are collected to satisfy the requirements of Community policies. The statistical needs expressed in the Rome and Paris treaties reflect the economic culture and policy requirements of their time. The treaties laid the groundwork for a legislative framework for data collection that centred on Member States, since comparisons between Member States were more important than comparisons between the Common Market and the rest of the world.

KEY WORDS: Response burden; Data use; Business statistics; Community regulations; Community information; Statistics and policy.

1. COMMUNITY POLICIES AND STATISTICAL NEEDS

In 1951, rebuilding the iron, steel and coal industry was a major concern. The emphasis was on integration. The Senior Authority collected statistical information directly from businesses, completely bypassing the Member States. The data were used both in fulfilling the aims of the European Coal and Steel Community (ECSC) treaty and in managing the current production and trade crises.

In the 1957 Treaty of Rome (Article 213), the collection of statistical data was dealt with quite differently. There was less emphasis on integration. The Community's statistical needs were still viewed as important, but the direct link between the central authority and individual enterprises was gone. The Commission asked Member States to collect the data. In this case, the subsidiarity principle was applied even before it was entrenched in Community documents.

EEC legislation called on Member States to establish a reasonably modest statistical system that would focus on the goods-producing industries. Services and trade were virtually ignored, despite their status as one of the four freedoms. Then came four major directives: 64/475/CEE on investment, 72/221/CEE on business structure statistics, 72/211/CEE on current conditions in the goods-producing industries, and 78/166/CEE on current conditions in the construction industry. Harmonization was carried out later, based on data collected and processed by Member States for their own purposes.

As the pace of Community integration increased, it became clear that the legislation on goods-producing industry statistics was inadequate to support the process. After the Single European Act calling for the creation of the Single Market was signed in 1986, some 300 directives were enacted to clear away the obstacles to the four freedoms of movement. Statistics were needed to evaluate the impact of the various measures that Member States took to make the Single Market a reality.

With the Single European Act's first steps toward economic and monetary union came the need to gauge the monetary and fiscal policies of Member States. Greater harmonization in the measurement of inflation, unemployment, foreign trade and production (the magic square) became essential.

When the sources of funds for the Community's budget were expanded to include a percentage of the VAT base and a percentage of GDP, Member States began demanding virtually complete harmonization of GDP concepts and definitions. For political reasons, it became increasingly important to highlight the advantages of belonging to the Community and to measure the returns of European solidarity. The aim of European statistics could be summed up in four words: comparable, complete, accurate and timely.

The Maastricht Treaty of 1991 established the European Economic and Monetary Union and set out the requirements that Member States would have to meet to join the Union. This entailed measuring economic convergence with unprecedented precision and, consequently, putting in place much stricter, more specific legislation regarding data collection. While a complete list of statistical legislation adopted since then would be too long and tedious to include in this paper, it is worth noting that systems of national accounts were made mandatory and were fully harmonized. The consumer prices measurement system and the interest rate policy monitoring system were also tightened considerably. Budget-related concepts were harmonized and given more rigorous definitions, and specific nomenclatures were formalized in regulations.

Since 1990 the following legal framework has been established for business statistics:

General Statistical Legislation

- Council Regulation No. 322/97 of February 17, 1997, on Community statistics.
- Commission Decision 97/281/CE of April 21, 1997, on EUROSTAT's role in the production of Community statistics.

[1] François de Geuser, Commission Européenne Eurostat, Bâtiment Jean Monnet, C5/57, Rue Acide de Gasperi, Luxembourg-Kirchberg, L-2920, LUXEMBOURG.

- Council Regulation (CEE) No. 2223/96 of June 25, 1996, on the European system of national and regional accounts in the Community.
- Council Regulation (EURATOM, CEE) No. 1588/90 of June 11, 1990, on the transmission of confidential statistical information to the Statistical Office of the European Communities.

Nomenclatures and Classifications

- Council Regulation (CEE) No. 3037 of October 9, 1990, and Commission Regulation (CEE) No. 761/93 of March 24, 1993, on the statistical classification of industries in the Community.
- Council Regulation (CEE) No. 3696/93 of October 29, 1993, on the statistical classification of products by industry (CAP) in the European Economic Community.

Statistical Tools

- Council Regulation (CEE) No. 696/93 of March 15, 1993, on the units of statistical observation and analysis to be used for the Community's goods-producing industries.
- Council Regulation (CEE) No. 2186/93 of July 22, 1993, on Community coordination in the development of business directories for statistical purposes.

Collection Regulations

- Council Regulation (CEE) No. 3924 of December 19, 1991, on the creation of a Community survey of industrial production.
- Council Regulation (CEE) No. 3330/91 of November 7, 1991, on statistics on merchandise trade between Member States.
- Council Regulation (CE, EURATOM) No. 58/97 of December 20, 1996, on structural statistics on businesses.
- Council draft regulation on current economic statistics (COM (97) 313 final).
- Council Regulation (CEE) No. 2744/95 of November 27, 1995, on wage structure and distribution statistics.

Accounting Directives

- Fourth Council Directive (July 25, 1978) based on Article 54, subparagraph 3g, of the Treaty, concerning the annual accounting statements of certain types of corporations.
- Commission recommendation of September 13, 1995, on taking into account the statistical nomenclature of industries in the European Communities in breaking net earnings down by industry.
- Seventh Council Directive (June 13, 1978) based on Article 54, subparagraph 3g, of the Treaty, concerning consolidated statements.

2. CONSEQUENCES OF INCREASED STATISTICAL NEEDS

The introduction of this completely new legal framework exerted heavy pressure on the national statistical systems and the businesses supplying the raw data.

2.1 Changes in Scope and Type of Data Required

For the national statistical systems, the effects could – and in some cases still can – be compared to those of a natural disaster. National collection nomenclatures were Europeanized. Business registers had to be completely overhauled and rebuilt. In some countries, business registers had to be created, and major amendments had to be made in national legislation. Changes in collection nomenclature led, in some instances, to changes in the classification of businesses for collective agreement purposes. For some countries, the introduction of new statistical units created many difficulties, though none of them were serious.

There was also a change in scope. Under the old system, only goods-producing businesses with 20 or more employees were surveyed. Now, all businesses, regardless of size, and all industries except farming and government had to be covered. For the National Statistical Institutes (NSIs), this meant that the number of businesses to be surveyed jumped from between 50,000 and 100,000 to between 200,000 and 4 million; for the Community as a whole, the number of units increased from 500,000 to about 16 million.

The data requirement changed as well. Whereas only the main accounting aggregates were needed before, now the volume/price distribution had to be computed. Productivity had to be measured not only in goods-producing industries but also in services; as a result, it was necessary to consider what productivity meant in the service industries. The concept of benchmarking had to be introduced to assess the competitiveness of businesses, industries and even Member States.

Information was needed about the "demography" of businesses in order to measure the impact of assistance policies for small and medium-sized businesses (SMEs). This "demography" would help define the prerequisites for SME creation and survival, and the job-creation effectiveness of SMEs.

2.2 Pressures on the Statistical System

Unfortunately, the added pressure on national statistical systems came at a very bad time for their budgets. Data collection costs soared just as national and European politicians were cutting the resources allocated to statistics.

The respondents (*i.e.*, businesses) perceived an increasing demand for statistical data. Certainly, the increased requirements due to Community policies led to a demand by the statistical system for a much broader range of information; to meet this demand, many businesses had to do some sophisticated processing on data from their internal accounting systems (*e.g.*, "end of pipe" investment, environmental investment, hours worked). Moreover, extended coverage meant surveying SMEs, which viewed the demand for statistical data as further government interference affecting their competitiveness. The extra burden on SMEs was also at odds with the administrative streamlining favoured by most European countries.

Finally, the quality of the statistical output was impaired by serious delays in publication, and businesses complained that they had nothing to gain by responding to surveys.

3. THE NEED TO EASE THE DATA COLLECTION BURDEN

We need to respond to this perceived increase in the burden on both businesses and the NSIs by finding ways to lighten the load. In another paper presented at this session, Mr. Machin described a number of specific steps that could be taken to measure and alleviate the burden.

I would simply like to say here that in the Commission's view, easing the collection burden is so important that it has proposed to the Council that the data collection regulations should contain a formal commitment to that goal, along with a statement setting out four means of achieving it:

- using existing information;
- using increasingly sophisticated statistical techniques;
- making survey response easier;
- explaining the uses and value of the information.

3.1 Using Existing Information

Article 6 of Council Regulation No. 58/97 of December 20, 1996, on structural statistics on businesses states that to reduce the response burden, national authorities and the Community authority have, in their respective jurisdictions and within the limits and conditions defined by each Member State and by the Commission, access to sources of administrative data concerning the areas of activity of their own public administrations, in so far as those data are necessary to meet the precision criteria set out in Article 7. In other words, existing data can be used for statistical purposes, which means that businesses will not be asked the same question several times by different administrations. It also means not only that the results of administrative surveys (taxation and customs forms) that collect personal information for the production of aggregate statistics can be used, but also that data collected for specific administrative purposes can be used for general information. In many Member States, constitutional rules will have to be amended to permit what some people refer to as information diversion. It will also be necessary to institute and guarantee confidentiality so that businesses can be sure that the statistics derived from the data cannot be used against them individually or to damage their competitive positions.

The fact that permission was given to use existing data suggests the possibility of linking different types of files relating to businesses, including statistical files, tax files (profits, VAT), social security files and chamber-of-commerce files. Approval would have to be sought, of course. The Council has adopted a regulation to coordinate the contents of business registers. Some Member States have to pass special legislation to permit the linking of files. The idea of having a single identifier for each business is by no means unanimously accepted in the Community. The drive to streamline and economize by using existing data must not blind us to the enormous difficulties that have to be overcome to obtain acceptable data quality. The various surveys and the various administrative files all have different purposes. Thresholds set for purely administrative reasons are hard to deal with in the statistical process. In some cases, populations are defined very differently. Nevertheless, through statistical research it should be possible to find solutions to these problems so that existing data can be used to lighten the response burden.

3.2 Using Statistical Techniques

The above-mentioned regulation also calls for the use of sophisticated statistical techniques and refers to one technique in particular: sampling. The program of symposium '97 on new directions in surveying and census-taking lists all the available options for simplifying and lightening the statistical burden. I will not go into further detail on this point, but I would like to point out that it must become standard practice to use sampling and other techniques such as small area estimation, profiling and panels.

3.3 Making Survey Response Easier

In the same Article 6 of Regulation No. 58/97, Member States and the Commission are invited to foster the increased use of electronic transmission and automated processing of data. The purpose is to make it easier for businesses to respond to surveys. It is not the same as simplification. The SERT project is intended to ease the reporting burden on businesses (especially SMEs) and governments by automating the collection of data and the dissemination of processed information to businesses.

SERT is a series of coherent, effective actions designed to phase in the two-way exchange of information between businesses and NSIs. The raw data are generated automatically by businesses or their agents (accountants, professional associations, *etc.*).

- Automated data collection speeds up processing, and faster processing means more timely statistics and more relevant analyses.
- In addition, automated collection is closer to the data sources, and the closer they are, the easier it is to prevent distortions due to differences in the way data is processed by the various businesses and the various countries.

The data are collected at the most detailed level rather than at aggregate levels. For businesses, most statistical information is present either in the accounting records or in the data used to generate those records. After all, a business's accounting records are simply a model, subject to certain rules, of the financial and capital consequences of decisions made by the business's management.

This brings us back to SERT's main purpose, which is to model the business statistics database (BISE) and enable the two-way exchange of information between businesses and data collectors. At the same time, SERT reduces the respondents' workload and costs, makes data collection more efficient, and speeds up the production of statistics, which in turn increases their value for analysis and economic forecasting.

For example, the use of tools such as RDRMES (Row Data Reporting Message) and GESMES (Generic Statistical Message) has speeded up the forwarding and processing of production (PRODCOM) and balance-of-payments (BOP) data.

A number of trials have been conducted to bring respondents and surveyors closer together, including EDICOM-IDEP for INTRASTAT transmission, and a project to centralize response to iron and steel surveys through the European Confederation of Iron and Steel Industries (EUROFER). A very promising study of private and public nomenclatures and the possibility of establishing highly automated links between accounting codes and statistical variables in the BISE context is under way in France, Belgium, Greece, Austria and Denmark. It will probably lead to development, in 1998 or 1999, of accounting software that will electronically transmit administrative and statistical questionnaires without user intervention.

There have been a number of pilot projects involving the transfer of raw data from businesses to government offices using EDIFACT RDRMES:

- collection of PRODCOM data in the United Kingdom;
- Project EDIVAT in Belgium. In this project, completed in 1996/1997, accounting firms used RDRMES to transmit VAT returns to the Belgian VAT administration on behalf of their clients. The project was so successful that it won the Interchange 1997 Award. Electronic filing will be officially accepted on January 1, 1998, as an alternative to submitting the paper form;
- Tourism SERT project in Greece. Cooperation between the Greek Statistical Institute, the Technical University of Athens, and Abacus, a publisher of hotel management software (52% of the market), led to the RDRMES modelling of the monthly and annual tourism surveys, the development of a telecommunications interface, and the full integration of those modules into the September 1997 version of the software (in an entirely transparent way).

3.4 Promoting the Use of the Statistics

Response burden is a rather subjective concept in that it consists of the difference between the costs of responding and the expected benefits (*cf.* Mr. Machin). Yet the fact remains that the lower the perceived return of useful information – it may even be nil if the production takes too long – the greater the perceived response burden. When current indicators of production are published nine months after the reference period, one can be forgiven for questioning their value. One might even go so far as to doubt the value of including them in the Treaties. Hence, in order to ensure both accurate responses and data reliability, it is a good idea to encourage businesses, especially SMEs, the most vocal opponents of surveys, to use statistics. Arranging for the return of information to SMEs, presenting studies and analyses on the uses that SMEs can make of statistics, showing how SMEs that do not use statistics can fail, and showing that statistics provide independent, scientific information will go a long way toward remedying statistics' negative image.

EUROSTAT's role is to act as an advocate for statistics both in academic debates and in public policy debates about the role of SMEs, job creation, and so on.

To conclude, I initially found it odd that EUROSTAT, which does not collect statistical data, was invited to present its views on data collection. I realized, however, that EUROSTAT has an important role to play in coordinating the various measures taken to lighten the response burden borne by businesses:

- preparing a best-practices guide showing how information from various sources can be used;
- making survey response easier by automating and streamlining it;
- making a systematic effort to test and improve data quality and thus increase user confidence;
- convince people that statistics are not a public-sector monopoly and that the challenge issued by users and private statistics agencies must be met.

PANEL DISCUSSION

Approaches to Innovation in Statistics Agencies

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

PANEL DISCUSSION ON MANAGING CHANGE IN STATISTICAL AGENCIES

W. McLennan[1]

1. INTRODUCTION

Over the last 10 years there has been massive change and innovation in all areas of the Australian Bureau of Statistics (ABS), which has resulted in increased productivity. Managing such change and innovation is, of course, dependent on the culture of the organisation, its management style and its ethos.

There has been strong and consistent input from methodologists as Susan Linacre, who besides leading our Methodology Division, has a corporate responsibility for methodology and its application right across the ABS. Methodologists have, for a long time, been encouraged to be involved in operations to give them a practical slant and to move within the ABS from and to the methodology area. As a result there is strong mutual trust and respect across the organisation, and the services of Susan's team are keenly sought.

A separate point, but one which is fundamental to managing change and innovation, is that people in managerial positions must both manage and lead; unfortunately this is not always so. This problem must be addressed otherwise there will be no change or innovation!

2. TYPES OF CHANGE/INNOVATION

Taking a line through ABS experience, there seems to be two main types of innovation. The first is the strategic or broad change, which is usually instigated, or agreed to, by the top of the organisation, perhaps even by the CEO. The second is tactical, which is often processed based and incremental. I will discuss each of these approaches in turn.

2.1 Strategic Change

As might be expected with any strategic change, a clear and well communicated vision for the future is essential. Any organisation needs to know where it wants to be in 5 to 10 years time, so that change can be aimed in this direction. From our experience, there seems to be two ways strategic change is implemented. The first, which I will call the "light many fires" approach, comes about by encouraging all managers to institute change which is consistent with the organisation's strategic directions. One needs to communicate to staff that it is OK to give it a go. In the ABS such change has a big and positive impact on the work program. However, not only do you have to encourage many fires to be lit, but you have to be prepared to put some of them out quickly if they look like getting out of control or burning the wrong thing! Line managers need to be vigilant.

The second, and perhaps most important way that strategic change comes about, is by what I would call major projects. These tend to be strategic risk taking projects usually agreed upon at the top level of the organisation, commonly by the CEO. These projects usually relate to a core area of business, and must address a real problem. It is essential that both management and staff be on-side with the development. Often such projects are expensive, and if not handled correctly, perhaps devastatingly so. For many such projects, it is often a gut feeling decision to go ahead; what the project is trying to achieve is often a "known good thing" but sometimes not well defined.

2.2 Information Warehouse

A significant current example of a major project in the ABS is the information warehouse, a topic which has been well discussed at this symposium. It was started to ensure that the provision of ABS data matched the expectation of users, and hence a corporate rather than a collection view of the issues was required. In essence it is a corporate repository, *i.e.*, database, for all ABS output data sets and all associated meta data. Its aim, from the users point-of-view, is to make ABS's data and meta data visible, accessible, relatable, reliable, understandable and media-independent.

The very hard development and loading work is now largely behind us, and we are beginning to make progress in using the Warehouse to achieve statistical integration objectives. Most ABS data is now loaded to and regularly updated in the warehouse, the majority of information requests are answered from it and most economic statistics publications are produced using warehouse data and facilities.

Unlike most other large change projects in the ABS, this one has always received the full support of senior management and the staff, even through all the tribulations the project has faced.

2.3 Tactical Change

Let me turn now to tactical change. Again there seems to be two approaches. First, although the ABS has not

[1] William McLennan, CBE, AM, Australian Bureau of Statistics, PO Box 10, Belconnen, Australia 2616; e-mail: bill.mclennan@abs.gov.au.

implemented Total Quality Management (TQM) across the organisation, there is a lot of TQM like activity which fosters incremental change. This is encouraged by senior management but proceeds largely without their involvement. In the ABS, methodology and technology have been powerful enablers of this type of innovation. I would like to comment briefly on two examples, the 1996 Population Census success, and our recent work to integrate our statistical collections of businesses.

2.4 Population Census

Over the years the Population Census has been a classic example of continual incremental change. This is particularly true of the 1996 Census, which will be recognised as one of the most efficient and accurate censuses run in Australia. The continuous improvement approach, which involved staff right down the line, was particularly effective in the single processing centre established in Sydney, where data from all forms were captured, mostly electronically, coding was undertaken and clean data records were produced. The net impacts were significant savings, better quality data, and the production of first data output, covering all items which were self-coded, 9 months after Census day, and the final outputs 15 months after Census day, which were significantly earlier than any previous Census.

2.5 Survey Integration

As is the case for most statistical offices, the business surveys conducted by the ABS were largely developed independently. As a consequence, there were important differences in the statistical methods used. Although standard statistical classifications were used, and their frameworks based on the ABS business register, these frameworks were selected at different points of time even when the reference period was common. Edit/imputation methods were different, non-response was treated differently, and importantly, the means of estimating for new businesses not yet on the business register and dealing with defunct units were not common. Staff of the Methodology Division drew our attention to this and the significant contribution it was making to the lack of coherence in the national accounts and macroeconomic statistics more generally. Subsequently, the ABS had adapted a fully integrated approach to all its business collections, including the methods used in these collections. For example, they are now selected off a common framework and all use the same method for estimating new business provisions, utilising information available through the tax system on the number of new businesses. The benefits to the national accounts are already apparent. Other users will also benefit from the greater cohesion of the business collections.

The second tactical approach involves looking at issues from an holistic point-of-view, perhaps more commonly known these days as the re-engineering approach. It is interesting to note, although I said earlier that technology helps foster incremental change, the increasing rate of change of technology does sometimes push us in the direction of changing things from the bottom up. Again I would like to give two examples, the redesign of our Integrated Register of Businesses and the development of an expert system for seasonal adjustment.

2.6 Business Register

The ABS built its current business register system in the early 1980s, and although the system is still operational, it is expensive to run and maintain and not very flexible. Also, as you can imagine, the user interfaces are not very friendly by current standards. It was time to re-engineer the system but the large cost of doing so was a major deterrent. The proposal to redevelop came as a consequence of a series of discussions between ABS technology staff and our major computer supplier, Fujitsu. We committed by an almost "act of faith" decision to a joint venture with Fujitsu using their new object oriented database, ODB2. This was a big risk, with both the ABS and Fujitsu sharing the risk, but we now have a system. Its performance is not yet at the level we require for our target of 50 concurrent users, although it does work well for 30 concurrent users. We are confident of getting there – it will result in an innovative system that is much more user friendly, flexible and considerably cheaper to maintain. However, it has not been all plain sailing.

2.7 Seasonal Adjustment

The development of SEASABS, a networked PC based seasonal adjustment package driven by an expert system, is a classic example of technological change forcing a broad review of a project. The ABS had such a system which was mainframe based, but as it was complex to use it was mainly driven by time series experts, and there were too few of them. The advent of networked PCs and a suitable expert system enabled a user friendly and robust system to be developed. The net result is that users in subject matter areas do most of the straightforward seasonal adjustment work, leaving the small number of time series experts to deal with the more complex problems. It has been an outstanding success.

2.8 Success Enablers

What then were the main factors which have contributed to the ABS's success with change management? I'm not certain of the answer, but the following factors, I believe, are relevant:

- competent, motivated people/highly skilled/multi skilled
- emphasis on training/development
- good project management, even though it could be better
- clear lines of responsibility
- good conceptual and technological infrastructure, and

- often the active involvement of methodologists.

Considering the other side of the coin, what are the factors which might inhibit success? Again I'm not sure but the following are candidates:

- statisticians generally are not attuned to responding rapidly;
- corporate decision making in the ABS takes time and effort;
- too often we try for the 100% solution, or conversely we rarely accept the 80% solution;
- with a large investment in infrastructure there is pressure (perhaps of the moral variety) to use it, *i.e.*, the sledgehammer to crack a nut syndrome.

3. CONCLUSION

In conclusion, the following points, I believe, were important to the ABS' success in managing much change and innovation over the last 10 years, and hopefully, they will continue to contribute positively in the future:

- ABS's corporate approach to decision making supports change and innovation;
- The traditional place of, and culture surrounding, methodologists in ABS does so as well;
- Risk taking is accepted (by and large), in that occasional failure is tolerated;
- There is strong corporate ownership and contribution to innovative activity.

CONTRIBUTED PAPERS

SESSION C-1

Imaging, Integration and Automation in Data Processing

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A REVIEW OF DOCUMENT IMAGING FOR THE 1996 CANADIAN CENSUS OF AGRICULTURE

C. Bradshaw[1] and J. Duggan

ABSTRACT

Interest in the benefits of document imaging has existed for the Canadian Census of Agriculture for over 15 years. Only recently, however, has the cost of the technology reached an affordable level for our purposes. For the 1996 program, document imaging was successfully implemented with some results that exceeded expectations. The introduction and use of imaging technology is examined from several different perspectives: planning, implementation, production, and evaluation. In conclusion, the paper examines the overall benefits of the imaging solution to the 1996 Canadian Census of Agriculture.

KEY WORDS: Document imaging; Scanners; Image retrieval and display; Census production.

1. INTRODUCTION

In prior Censuses of Agriculture, a questionnaire library was maintained to allow editors and analysts access to individual documents during production and evaluation of the survey data. The library staff fluctuated between 5 and 15 employees over a 40-week period of receiving and filing questionnaires, accepting request lists for specific questionnaires, pulling these, and later filing them away again. Besides being a large resource to maintain, appropriate control records had to be kept so the location of a questionnaire could always be traced.

Imaging was seen as a solution to the primary problem of making questionnaires immediately accessible to all the processes that required the documents. In agriculture, there are large farming operations whose contribution to many data variables is significant. The questionnaires for these operations were frequently in demand, often by many analysts simultaneously. Photocopying was discouraged because of strict Statistics Canada confidentiality requirements. For the analyst, the process was often inefficient and frustrating. Documents selected for review required a written request and a wait for retrieval. These delays could be extended when requests involved missing questionnaires or those on waiting lists; documents would then be received sporadically.

2. PLANNING

A Statistics Canada (STC) project team was set up early in January 1994 with subject matter and processing specialists from the Census of Agriculture Section and representatives from systems development and methodology. The goals of the team were to investigate methods and procedures for reception, preparation, imaging (scanning, retrieval, and display) and storage of the 1996 Census of Agriculture questionnaires.

The major milestones of the research process involved two imaging pilot tests. The first assessed two scanners, a Fujitsu M3099A and a Bell & Howell Model 6338, in December 1994. The second test used a Kodak 500 D scanner and was also completed on-site in July 1995. The pilot tests were intended to answer many questions while promoting experience with the technology and the hardware.

Answers to questions about productivity, reliability, and the number of scanners needed were priorities. Concerns over the risk of breakdown, the initial costs, and the salvage value were also raised. Requirements had to be determined for image quality and storage, and these needed to be tested with the scanners and display software. From their research, the project team identified basic requirements for the major components (scanner, scanner management software, and image retrieval and display software) of the 1996 Census of Agriculture imaging system.

Scanner requirements covered: production volumes, duplex scanning, rated speed, resolution, barcode recognition, paper size, thickness and colour, standards for the feeding mechanism, hopper capacities and functionality, ease of operation and maintenance, calibration, and documentation. Quality standards were set to limit the skew of documents, double feeds, and barcode recognition errors. Technical requirements outlined the image type, file compression, PC interface cards, and available scanner settings. Lastly, warranty and technical support issues, including training and software upgrades, were described.

Scanner Management Software requirements involved: output characteristics of the images, record layouts, barcode translation, image enhancement tools, indexing of images into directories and exception directories, support for duplex scanning, audit files, and production statistics reports. The PC support requirements demanded a Windows application with a graphical user interface, server and local area network compatibility, and output file size/storage restrictions.

[1] Claire Bradshaw, Agriculture Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

The Image Retrieval and Display Software required: administrator-controlled security restrictions on accessing and printing images, retrieval of the 16-page questionnaire with concurrent access, annotation capabilities, on-line help, and image display features (*e.g.*, multi-page display, zoom, and panning). Additionally, images were to be delivered to the desktop within 5 seconds – and each subsequent page within 2 seconds – for an environment of 50 concurrent users and 100 total users. The system had to be launched from within an ORACLE*Forms application program interface on a UNIX server running LanManager-X. Specifications for retrieving scanned images from optical disk drives, juke-boxes and magnetic storage were also necessary.

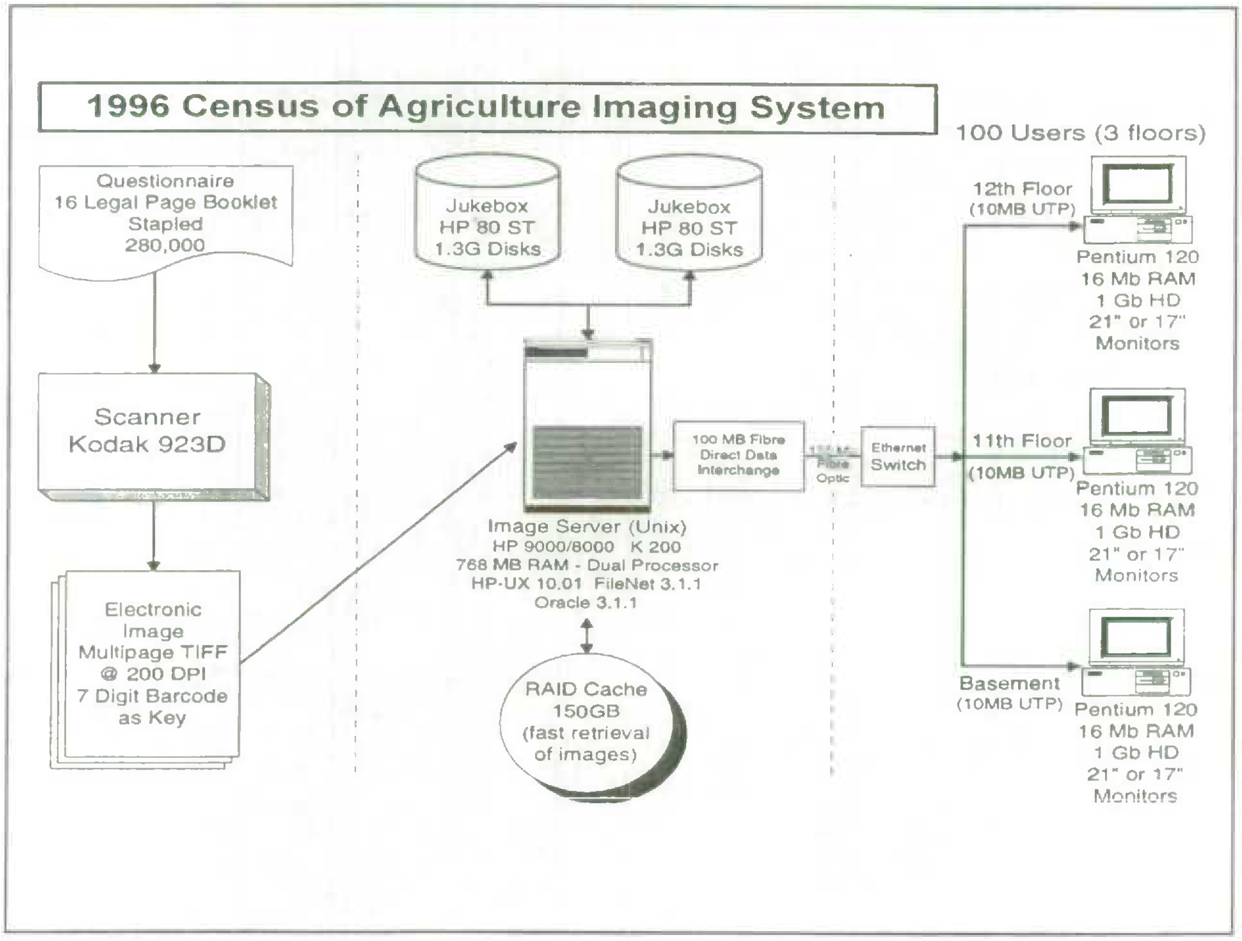
Major purchases by the Canadian government begin with a Request for Proposal (RFP) which invites bidders to tender for a contact. The RFP for an imaging solution was developed over a two-month period and was posted on the Open Bidding System in October of 1995 for two months, as required. Two qualifying bids were received, and the low bidder satisfactorily demonstrated compliance in a technical evaluation that was completed on February 9, 1996. The agreed-to contract specified that all hardware and software would be delivered by March 25, 1996.

3. IMPLEMENTATION

After the delivery of the imaging hardware and software, a system that integrated with the Census of Agriculture database had to ready for production for July 1, 1996.

The intervening three months were spent in preparation; implementing the solution and conducting thorough testing and re-testing. A team from the contractor spent two weeks at the end of April on-site to install and set-up the imaging system. Their work included the set-up of the delivered components, training of STC technical personnel on the software and its installation, and customization of their software to allow its integration with the Census of Agriculture database processing. The company was also required to provide on-site support for a further two-month period. The following chart documents the architecture for the imaging system.

The integration of the imaging system with the Census of Agriculture system required four customized processes. First, an index process to update the scanning software image identifier and link the image to the Census questionnaire barcode identifier was required. Second, a commit process to transfer the scanned questionnaire image to the image database was built. Then, an update process notified



the Census of Agriculture database of the new questionnaire image on the Image database. This third process also annotated the barcode identifier from the first page to all subsequent pages of the questionnaire image. This was done so that any one page could be viewed (or printed for use in one exception process) individually while still ensuring a correct identification. A fourth process, display, was created to permit pre-specified pages of one or more document images to be automatically displayed on the user's workstation upon retrieval.

4. PRODUCTION

The full production period for the Census of Agriculture starts about one month after Census Day, and is completed 11 months later. In 1996, the imaging of the 280,000 questionnaires was completed over a 14-week period. Since the data capture operations preceded the imaging operation, the production schedule was timed to keep pace with the flow of documents from regional data capture sites and to ensure there were enough images in the database to support the editing operations that followed.

Questionnaires were groomed, labeled with a barcode, and batched in an initial processing stage at five Statistics Canada Regional Office locations. The data were captured and then transmitted to Ottawa for loading to the Agriculture database. The physical batches of captured questionnaires were sent by priority post to Ottawa and were typically received at the imaging centre from one to three days later. The batches were checked-in with a barcode reader and verified against the database record for the batch. The stapled spines were cut from the booklet with an automatic electric cutter to prepare for the scanning operation. Cut questionnaires were replaced in the batch envelopes and moved to the scanning bay.

The batches were scanned with a Kodak Imagelink Scanner 923D equipped with a barcode reader. The legal sized pages were scanned at 200 dpi in the duplex setting and were fed into the scanner in portrait mode, leading with the top edge of the questionnaire. The final average file size for one 16-page questionnaire was 643,000 bytes. The speed of the scanner averaged 104 pages (6.5 questionnaires) per minute during the production period. This included a 20 to 40 second software initialization sequence for each batch, handling time, and periodic maintenance. Under ideal conditions, the imaging production reached 128 image pages (8 questionnaires) per minute, but this level was not sustainable for more than three consecutive hours. At the peak of production, 100,800 image pages (6,300 questionnaires) were imaged in a 15-hour period. The scanner regularly operated on weekdays over a 9-hour period with an evening or weekend shift added periodically when production schedules warranted.

The rate of double feeds, where two or more pages are pulled into the scanner at the same time, was less than 1%, (2,700) of the questionnaires. The flow of imaging production was not interrupted to locate and rescan these questionnaires; instead, the system created an exception log of any questionnaire images that did not have exactly 16 pages. The system also separately identified questionnaires from the control file for which no images existed. These errors and omissions were corrected immediately if the document image was requested, but otherwise were attended to as time permitted later in production. Users identified a few low quality images with the aid of a "Rescan" button built into the image display software. The rescan requests were also treated with priority and the documents were rescanned as the requests were received.

The scanner was only forced to stop operating on a small number of occasions. Out of more than 700 hours of operation, only about 37 hours (5%) were lost due to technical difficulty with the scanner. Of these 37 hours, 24 (65%) were attributable to a problem in the imaging centre with an electrical supply that had not been correctly wired (dedicated). The remaining downtime related to an adjustment to the barcode reader at four weeks of production, a sensor circuit board replacement two weeks later, and a front lamp adjustment with only a week left in production.

A few start-up problems with image retrieval and display were encountered at the beginning of the process. The imaging software and the production programs were running on the same server that stored the Oracle database. As processing geared up, the entire system slowed and the server could not handle the volume of images delivered by the scanning process. Better scheduling of background jobs and rewriting of some of the more intensive programs resolved the problems.

5. EVALUATION

In discussing how the electronic images were used in the processing of the Census of Agriculture, it may be helpful to consider the work as being done in two separate stages. The first stage, production, can consist of the various coding and editing functions necessary to prepare estimates from the captured data. The second stage, analysis, validates data through investigations at both the macro and micro levels.

In production, a series of processes is executed in a generally linear fashion as data from each questionnaire follows the flow of processing. In previous censuses, the passing of physical documents from one operation to another would have marked these stages. In the image-enabled environment the conceptual links still existed, but the questionnaires remained in the library. An automated programme examined each questionnaire record on the database to identify records in need of correction or acknowledgment. These selected records would be internally batched by the system and presented to operators for 'manual' resolution through an on-line processing facility. These 'screens' varied according to function but each was designed with an icon to allow the appropriate page of the image for the record in error to be displayed to the side.

The analysts began their work of assessing and ensuring the quality of the data by examining trends and patterns in sub-national and other aggregate levels of estimates. Agricultural survey estimates and related sources such as administrative data and lists were supplied. In addition, the analysts were able to view Census estimates at several stages of processing to note any potential problems created by production. Investigation naturally involved micro level analysis, checking that respondent data supported the observed macro-level trends. Analysts made heavy use of the images, when it was necessary to view the initial responses to the questionnaire. This low-level detail is of critical importance to the census as it provides input to the central frame for a number of smaller agricultural surveys and also supports the logical consistency of data for small areas.

6. OVERALL BENEFITS

The Census of Agriculture realized many benefits from the implementation of imaging. These benefits were identified by evaluating program objectives concerning the quality and timeliness of the work against the net cost of the imaging solution.

Use of the electronic images of questionnaires resulted in a great reduction in the amount of paper handling that was required. Individual tasks were faster and easier because the images were available on-line beside the other tools that were needed to perform the work. Navigating, or moving from page to page and from document to document, was more exact and took less time. Many documents could be open simultaneously without cluttering workspaces and could be 'put away' as easily as they were retrieved. The automated preparation of batches simplified the operations; images eliminated the need to continually separate and sort questionnaires. A tighter control of the documents, the processing, and security resulted.

These gains were achieved at the same time that the role of the questionnaire library was reduced. Also, the need to maintain a separate operation to control and co-ordinate the delivery of questionnaires to individual processes and analysts was eliminated. The fast and reliable availability of images to many users at one time was made possible in an environment that allowed automatic tracking and feedback concerning the usage of the images.

The quality of the analysis was improved on two fronts. Firstly, the speed and reliability of image retrieval allowed more images to be looked at in a shorter amount of time. In all of 1996 production and analysis, 48% of the images were viewed with an average of 1.7 hits per image, a level of 82%. A comparison to 1991 levels involves adjusting for changes in the order of processing (designed to take advantage of the availability of images). The adjusted level of requests in 1996 was 45%, much higher than the 25% of all questionnaires requested in the previous Census. This increase reflects the increases in both production and analysis. Secondly, the quality of the analysis itself was improved. Factors included the availability of image-enhancement tools, making some images easier to read than the originals, and the ability to annotate comments directly on top of the questionnaire image, a valuable form of communication among analysts. *Ad hoc* analysis was further assisted and improved by the prompt and ready display of images. An analyst could pursue a line of investigation promptly, without having to document an inquiry, hoping to recover her train-of-thought after a lengthy wait for all required documents.

Unlike analysis, where time saved was reinvested in more analysis, the production process benefitted from visible and measurable savings in time. In one particular example, the clerical edit process was completed more than six weeks earlier than the 22 weeks projected based on previous Censuses. Also, a couple of contingency operations, a new follow-up survey and a second collection using Computer-Assisted Telephone Interviewing, were added into the production cycle without compromising the usual release date (one year after Census day). These successes would not have been possible without electronic imaging.

Another unexpected quality benefit appeared during processing. Employees who had worked on the previous survey in both the analytical and production areas were more satisfied with their jobs. The work was less menial, improved the technical skills of staff – a significant benefit to temporary workers – and contributed to a sense of accomplishment and satisfaction,

In terms of cost, the benefits of imaging did not reflect savings. Imaging required a large capital outlay to purchase the system. Imaging continues to cost $72,000 per year for maintenance, license fees, and system upgrades. The system also needs trained technical support. The start-up costs of the 1996 Imaging System were about $820,000 or $8,200 per user. This represented 3.2% of the total seven-year budget cycle for the Census of Agriculture. Off-setting these expenditures were direct savings of $450,000; identified from the elimination of the 1991 library and archiving operations, and savings realized in the clerical edit process. Additionally, cost recovery imaging contracts from other areas of Statistics Canada have helped pay the ongoing costs of the system. The imaging system is also currently being used for research and testing of an automated capture strategy for the 2001 Census of Agriculture.

REFERENCES

Duggan, J. (1996). Electronic Imaging in Support of the 1996 Census of Agriculture. Presented at the International Conference on Computer-assisted Survey Information Collection.

Green, I. (1995). Questionnaire Imaging Pilot Project. Internal Report, Statistics Canada.

Neily, L. (1997). Scanning Procedures, Problems & Recommendations. Internal Report, Statistics Canada.

Young, P. (1995). 1996 CEAG Middle System Image Integration Design Specifications. Internal Report, Statistics Canada.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

OPTICAL CHARACTER RECOGNITION – A BETTER WAY TO CAPTURE BASIC 2001 CENSUS DATA

K. Roberts and E. St. John[1]

ABSTRACT

The Census of Population Processing project is currently evaluating the use of Optical Character Recognition (OCR) technology for the 2001 Census, as part of an ongoing effort to improve timeliness and data quality while reducing costs. Research to date has established benchmarks for comparison between traditional key entry and OCR based data capture. Current research is concentrating on defining an optimal design for systems and operations. Future research will assess the feasibility of establishing an OCR based system for the data capture of the 2A (short form) census questionnaires in the 2001 Census.

KEY WORDS: Scanning; Imaging; OCR; OMR; Data capture.

1. BACKGROUND

Since 1981, Census of Population questionnaires have been key entered, with approximately 1,500 operators keying in 4 billion keystrokes of data in five months (July - November). Although efficient, the key entry infrastructure required is large and used for a relatively short time.

In consideration of the above, and in an effort to keep abreast of new technologies, the Census of Population is investigating *Optical Character Recognition* (OCR) technology to capture the short form (2A) responses in 2001. Essentially, this technology involves taking a digital picture or image of a page and then extracting respondent information using data recognition software. Currently, this technology is being used by Statistics Canada as well as by non-government institutions for forms processing.

The long-term goal is to improve timeliness and data quality while reducing costs. Furthermore, the Census will position itself for future processing activities by using current hardware/software.

2. PROGRESS TO DATE

A significant amount of effort has been invested by various census processing staff since April 1995 in researching the potential of OCR for the next census.

Preliminary discussions were held with the United States Bureau of the Census (USBC) concerning the use of OCR/scanning technology in the capture of their 2000 Census. Subsequently, the Census of Population project acquired an OCR system, similar to the configuration being used by the USBC, for preliminary research and testing. A research team was also established, with members from Census Operations Division, System Development Division, Methodology and Subject Matter divisions. A number of issues were identified for investigation, including the extent of pre-capture preparation for booklet questionnaires, recognition of numeric/alphabetic characters and quality control of the imaging process.

A baseline test was conducted to conclude the research done on this system. The test objectives were:

1. To create a base of measures that will be used as a reference base for cost/benefit evaluations by the 2001 data capture research project.
2. To elicit feedback on the use of images by keying operators.
3. To assist in identifying data capture research issues.
4. To assist in identifying imaging system requirements.
5. To provide input into the decision making process for data capture technology.

Additionally, this baseline test used actual 1991 Census of Population questionnaires, which were designed for traditional key-entry data capture (*i.e.*, not enhanced for OCR/OMR).

The baseline test found that marks were recognized correctly up to 99% of the time (Optical Mark Recognition), and handwritten numerics were recognized correctly 29% of the time (Optical Character Recognition). The numeric recognition rate was below expectations, but it is felt that if the short form (2A) is enhanced for OCR (*e.g.*, establishing recognition points, using dropout colours) that this rate will improve significantly.

In 1996, the Census of Agriculture implemented a successful Imaging and Document Retrieval system for their census, which provided them with valuable experience on the use of Imaging technology in a production environment (*e.g.*, document preparation, scanning, and image management). In order to build on the expertise acquired by Census of Agriculture and to minimize the cost to both projects, it was decided to establish a joint approach to

[1] Kevin Roberts and Eric St. John, Census Operations Division, Statistics Canada, 4-C8 Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. e-mail: robekev@statcan.ca.

2001 research. This joint approach allows Census of Population to make use of the equipment and expertise acquired by Census of Agriculture in 1996 and provides Agriculture with a new research base on which to build their systems for 2001.

The research team has developed key components of an OCR operation which will be evaluated by capturing (through OCR/OMR technology) the respondent data from a sample of the 1997 Census Pre-test questionnaires. Data recognition software, data repair interfaces, workflow software, and system architecture are some of the system aspects being evaluated. The use of scanners, questionnaire design, dropout colours, operational procedures, and image storage are also being investigated.

3. ON THE HORIZON

The upcoming 2001 National Census Test (NCT) in the fall of 1998 provides an ideal opportunity for the evaluation of the processes proposed for the 2001 census, as the collected respondent information can be used to test imaging/OCR technology. The 2A, or short form, census questionnaire has been chosen for OCR/scanning because the majority of fields are check-off boxes, with a few constrained numeric fields. The research team is working closely with the Questionnaire Design team, Subject Matter and Methodology to ensure that the 2A questionnaire produced satisfies both NCT and OCR testing requirements.

The 2A questionnaire will be processed through traditional (key entry) and OCR capture systems, which will provide a "truth" file to assess the accuracy of this new capture methodology. This will also allow comparisons to be made between OCR/OMR and key entry data quality.

Processing the 2A questionnaires through both traditional and prototype systems (manual and automated), comparisons between the two systems can be made. This will allow the research team to realistically evaluate the proposed systems and to create a more effective process for 2001 (Figure 1).

Another key issue being explored by the Census of Population is the investigation of various ways to reduce the census infrastructure, from both an operational and system perspective. For example, system infrastructure can be reduced by designing a system architecture that allows requirements to be met for a minimum cost, while being open and flexible enough to adapt to new realities (*e.g.*, system hardware may change as the census approaches as the cost of this equipment is very fluid).

With a project of this magnitude there will always be risks, but two key issues that must be dealt with by the Census of Population are:

- Creeping expectations. Our goal is to establish a clear and precise set of requirements and to concentrate our resources on fulfilling these requirements. One of the risks of new technology is the desire to make it "all things for all people", which invariably satisfies no one.
- Confidentiality. Improvements in the flow of data within the data capture project also means increased access to the data. We will need to ensure that the proposed changes do not compromise the confidentiality of the census questionnaires and data.

4. 2001 CENSUS – PROCESS FLOW FOR IMAGING OF 2A QUESTIONNAIRES

4.1 Process Descriptions

Check-in:

Questionnaires will be registered automatically using barcodes.

Document Preparation:

Questionnaires will be sorted, information transcribed (if necessary), completed and non-completed sides of the 2A separated, batches created and the batches razor cut (to separate pages and ensure consistent sizing.

Scanning:

Questionnaire sheets will be fed through a high-speed production scanner in batches, with the scanner imaging both sides of each sheet simultaneously.

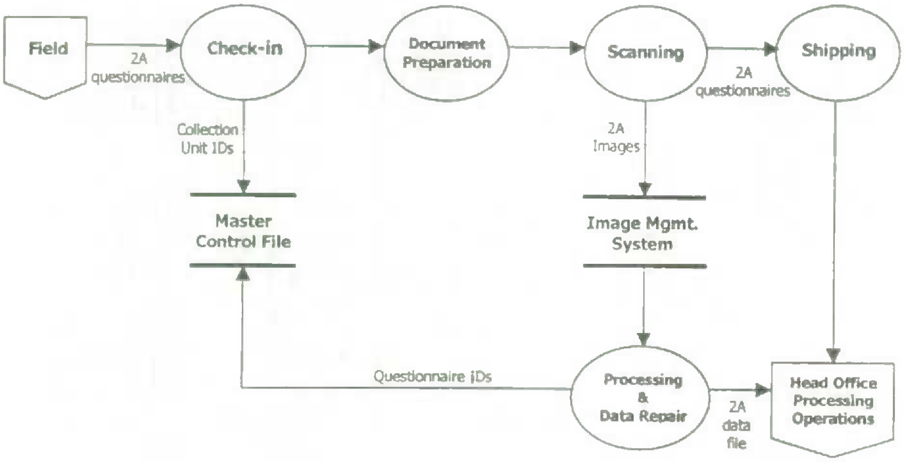


Figure 1

Processing and Data Repair:

The Processing and Data Repair Operation is responsible for the capture of data from the questionnaire images through the use of data recognition software, as well as the verification/correction of data that is not recognized.

Shipping (and Receiving):

The Shipping and Receiving operation is responsible for the physical receipt and shipment of questionnaires to and from the office. Additionally, this operation is responsible for ensuring that questionnaires are successfully scanned prior to shipment to HOP (via status reports).

5. IMPACT ASSESSMENT

Changes in technology invariably impact many aspects of an organization to some degree. OCR is no exception as far as Census Processing is concerned. This section highlights several (but likely not all) potential impacts on 2001 census operations due to the scanning of 2A short form questionnaires.

Changes in Process Flow

- More document handling, as pages must be separated prior to scanning.
- Separate process flows for the capture of the short forms (2A) which will be processed centrally and the long forms (2B) which will be processed regionally.

Questionnaire Design

- The 2A questionnaire will be enhanced to meet OCR requirements.

Equipment Requirements

- Servers (*e.g.*, data recognition, workflow, image storage) and laser disk jukeboxes required.
- Workstations needed for review and repair of recognized data.
- Fewer dummy terminals required for key entry data capture.
- Less mainframe capacity required.

Human Resources Impact

- Reduction in keying staff due to OCR/OMR of 2As.
- New technical support specialists must be available to maintain data capture software/hardware.
- Increased and more diverse training required for scanning operations.

Processing Space Needs

- Net impact will be a reduction in space requirements (approx. 20%).

6. CONCLUSION

Based on the direction STC is going with its other surveys, and considering the research results to date including projected benefits, the 2001 Census of Population will be capturing its short form (2A) responses using OCR technology instead of the traditional key entry approach. The immediate payoffs will include faster 2A OMR capture processing and keeping pace with new technology. Long term benefits include potential for processing cost savings for the census, satisfying archival requirements, use in field operations to support respondent follow-up, capture of the long form (2B), and improving quality. The risks include higher short-term costs for additional equipment purchases, staff training, *etc*.

The census must continually seek ways of improving and streamlining its processing techniques, while maintaining data quality.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

AN INTEGRATED APPROACH TO SURVEY DEVELOPMENT, PROCESSING AND DOCUMENTATION

L. Hunter and J. Ladds[1]

ABSTRACT

In Special Surveys Division, we have developed an integrated approach to Question Design, Data Processing and Survey Documentation through the use of information stored in relational databases. This approach was the result of efforts within the division to reduce the costs, time and errors associated with cost-recovery surveys. Information stored in the databases is accessed through an interface; commonly used functions are provided; and standardized outputs can be produced. There are two areas which are documented in these databases. The first is the set of variables used in the survey (question text, response categories, specifications for edits and derived variables and attributes such as data type, length and comments about data quality). From this database we can generate survey questionnaires, CAI specifications (currently for Interviewer, CASES or XSURV software), spreadsheets for testing scenarios, database structures or record layouts for the survey data files, extended codebooks, as well as SAS and SPSS cards for users to read the microdata files. The second area of documentation is the steps involved in the processing of the survey (objective, input file names, output file names, name of program, history of date/time program was run). From this database, we can control the steps of the processing and ensure that programs are run in their correct sequence and with the most up-to-date version of both programs and input files. We can also analyze the historical information from previous runs of the processing to identify bottlenecks and other problems that should be fixed for future iterations of the same or similar surveys.

KEY WORDS: Documentation; Metadata; Database; Code generation.

1. BACKGROUND

The Survey Documentation System (SDS) is a relational *database*[2] approach to organizing the documentation associated with the development, processing and dissemination of a survey. This approach was the result of efforts within Special Surveys Division to reduce the costs, time and errors associated with cost-recovery surveys. Four problems were identified as sources of errors in many surveys:

1. Re-typing survey questions and responses for questionnaires, CAI applications and codebooks; a great deal of the information is common to all (question wording, response category wording), although each has some unique elements.
2. The existence of multiple versions of data files. This has occurred because of updating the programs which produce those files without documenting anywhere that the output file name has been changed; old files get used as input to subsequent steps which creates cascading errors.
3. Manually editing the data without fully documenting what was done. Sometimes, project teams have decided to "fix" problem data manually rather than in a program; if anything happens to the resulting file, the fixes are difficult (and time-consuming) to reproduce.
4. Discrepancies between processing specifications (what you want) and reality (what was really done).

SDS was designed to eliminate these problems by providing survey teams with an appropriate structure and working environment. In particular, SDS uses relational databases to store information about the survey questions and variables, and information about the steps in the survey processing. The databases are accessed by users through a set of forms which allow them to view/edit the information, perform generic functions to change the information, or run routines which output information in a standardized format.

The SDS is composed of two main applications: DEVSURV for developing and documenting the survey variables and PROSURV for designing and documenting the steps in the survey processing. Each application has its own interface for accessing the underlying database. The term "interface" is used here to denote a series of related forms which access an underlying database. These applications function independently; however, their real power comes from using both together.

Both applications were developed by prototyping a set of tables, along with the interfaces to access those tables. The prototype systems were tested by project teams working on actual surveys. Their feedback was used in making modifications to be implemented in the full version. The modifications centered on (1) adding more fields to the tables; and (2) providing generalized functions or programs which use the documentation to automate the more common tasks.

SDS has been used for about 20 surveys since it was added as a utility on the division's network about two years

[1] Lecily Hunter and John Ladds, Special Surveys Division, Statistics Canada, 5-D5 Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

[2] A database is a collection of related tables (or files) which are associated with each other through common fields.

ago. The original objectives for the system have been met; the approach has allowed survey teams to enter information into the database and then use that information not only in the ways in which the objectives originally intended but in a wide variety of new ways as well. We have been able to reduce the time required to develop CAI applications; we can respond quickly to requests to produce non-standard reports; and we can analyze the history of the survey processing to identify recurring problems which can then be corrected for the next iteration of the survey.

2. DEVSURV

The Survey Development application (DEVSURV) supports the development of survey questions and the documentation of all information related to survey variables. Each piece of information is represented as a field in the database. Various combinations of these fields can then be retrieved from the database and formatted in different ways to produce a variety of useful outputs. DEVSURV ensures that all the characteristics associated with survey variables are kept up to date in a central location. Outputs are generated from a common set of programs ensuring consistency across surveys. Specifications entered in DEVSURV are used directly to generate code in the programming languages commonly used in the division. This puts the onus on the specifier rather than the programmer to correctly define what is needed, reducing problems of mis-communication between specifier and programmer.

The current version of DEVSURV encapsulates virtually all variable-level information which is required by most of the surveys done in Special Surveys Division. However, some of the more complex surveys have specific needs which are not currently supported. During the testing and development of DEVSURV, users were constantly coming up with new ways in which the information stored in the database could be used. These new ideas were incorporated whenever feasible. However, it was recognized that development would never end if all requests were to be incorporated. Instead, the decision was made to design DEVSURV in such a way that users could modify the standard system to meet non-standard needs. This was accomplished by providing the "hooks" for running user-designed forms and programs through the regular DEVSURV interface, as well as permitting project teams to add new fields to the database tables. Users are also encouraged to report on these modifications so they can be added to the list of future requirements.

Another way in which DEVSURV was designed for flexibility was to allow project teams to begin using it at different stages of the survey cycle. Teams which have already developed a questionnaire using a word processing software can import their questions into DEVSURV and then add more information to the survey variables, and use the tools for processing and dissemination. Completed or nearly completed surveys can be imported into DEVSURV to use the tools to produce dissemination outputs.

Some of the outputs produced through DEVSURV are:

1. Survey questionnaires.
2. CAI specifications (currently for Interviewer, CASES or XSURV software).
3. Spreadsheets for testing scenarios.
4. Database structures or record layouts for the survey data files.
5. Extended codebooks (with weighted and unweighted frequency counts).
6. SAS and SPSS cards.

Some of the functions built into DEVSURV are:

1. Validation routines to ensure that variables descriptions are consistent (*e.g.*, Comparing variable length to response category codes).
2. Calculation of start and end positions for each variable on the record layout.
3. Creation of de-strung "mark all that apply" questions.
4. Conversion of "don't know" and "refused" response codes from collection into the standard set of codes used for all surveys; adding codes for "valid skip" and "not stated".

Some of the benefits provided to the project team are:

1. Less time is required to produce reports which present the questions and variables in various formats.
2. Edits and derived variable specifications are stored directly in the database, making them easy to locate. In many cases, the program code can be generated directly from these specifications, thus reducing errors, saving time and ensuring that the specifications reflect what was actually done in the processing.
3. All the information about the survey questions and variables is stored in a central location and all updates to this information are made centrally. This makes retrieval of information faster and reduces the possibility of "version" errors.

3. PROSURV

The Survey Processing application (PROSURV) allows the project team to fully document all the steps involved in the survey processing. For each processing step, the team specifies the files to be used and the program(s) to be run, as well as describing the purpose of the step. Once this information has been defined, PROSURV will allow the team to actually run the processing programs. When the programs are run, information about the processing history is automatically recorded: the date/time the step was run, the time required to run the program, the reason for running/repeating the step, and the version of the program which was run.

PROSURV ensures that all the information needed to successfully run each step is kept up to date in a central location; each step of the processing is fully documented; and the entire processing can be repeated without manual "fixing". It also allows the programmers to remove themselves from the actual running of the programs; the programmers write the programs and handle any corrections, while the processing reps run the programs, check the results and update program specifications when necessary. This frees up the time that programmers often spend on "menial" tasks and allows them to concentrate on the more challenging aspects of the processing such as the data handling associated with complex and longitudinal surveys.

Some of the function provided by PROSURV:

1. Running the program(s) associated with a processing step.
2. Running frequency counts on the output file(s).
3. Running comparison counts, linking frequency counts from before and after the processing step was run.

Some of the benefits provided to the project team are:

1. Linking the processing documentation to the actual running of the programs discourages the use of non-programmed data manipulation, thereby reducing errors and problems associated with non-repeatable editing.
2. The historical information which is logged every time a step is run is a valuable aid to identifying problems in the design of the processing steps. Every time a step is run (or re-run), the user is asked to provide an explanation (*i.e.*, what went wrong the last time the program was run). For repeated surveys especially, this information is valuable in identifying where problems occurred and what was done to solve them – there is hard data on which to base decisions rather than relying on memory.
3. By writing programs which read file names directly from the documentation database rather than hard-coding them into the program, the correct version of a file is always used. Name changes need only be made in a single place. This reduces errors caused by old versions of data files being used as input for a subsequent processing step.
4. The learning curve for new project team members is shorter because the information about how the processing is done is well documented and easily accessible.
5. Processing multiple surveys concurrently is made easier because all the information needed to run the processing for a given survey is bundled together; the PROSURV interface makes running the programs a simple matter of pushing a button.

4. CONCLUSIONS AND FUTURE DIRECTIONS

Storing survey documentation in relational databases has allowed us to create an environment in which information is entered, updated and stored centrally and is available not only to the SDS applications, but also to any other program which could benefit from using it.

The SDS approach encourages the use of standardized tools for tasks which are common to all (or many) surveys. Although only a few of these standardized tools have been developed, many more are under consideration for future development and implementation. Outputs and reports have the same "look and feel" across surveys, making it easier for the dissemination staff as well as data users to find the information they want.

Survey teams have realized time savings during question development and data processing. The extent of these savings depends directly on the extent to which the team makes use of SDS tools and capabilities. As more project teams become familiar with the information available in the databases as well as how to use that information, more savings will be realized.

The SDS environment encourages the creativity of project team members. By identifying new fields to add to the databases and ways in which those fields can be used, the usefulness of this approach will be expanding for some time to come. Looking for actions to automate or ways to improve our current approaches can be very rewarding to experienced staff.

Over the next year, the SDS applications will be enhanced to include suggestions gathered from users in Special Surveys Division over the last two years. In particular, we will be adding more fields to the DEVSURV database and improving the interface so that it will be better able to handle longitudinal and complex surveys (for example, documenting variations in question wording between collection cycles and identifying different units of collection/analysis). Special Surveys Division will also be developing a set of generic processing tools for handling the common steps in survey processing; these tools will be directly accessible through PROSURV.

Although we have already derived many benefits from setting up survey documentation in a relational database, we believe that we have only begun to scratch the surface. Many more improvements will be possible as we analyze how we are currently doing our work, identifying where information already exists (or could exist) in the documentation database and then adjusting our usual methods to make use of that information rather than repeating or ignoring it.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

GENERALIZED TOOLS FOR THE AUTOMATIC GENERATION OF EDITING AND CODING APPLICATIONS FOR STATISTICAL SURVEY DATA

R.M. Hanono[1] and D.M.R. Barbosa

ABSTRACT

Following the new trends in automation, a new data processing strategy for the editing and coding phases of statistical survey data is being used at IBGE – Brazilian Institute of Geography and Statistics. It provides decentralization, portability, independence and integration, bringing the subject – matter specialists next to the process. For this purpose we developed at IBGE the automatic generators, "CRIPTAX" and "SISCOD", for editing and coding applications, respectively. The tools features and their usage at IBGE will be described in this paper.

KEY WORDS: Editing; Coding; Generators.

1. IBGE - BRAZILIAN INSTITUTE OF GEOGRAPHY AND STATISTICS

IBGE is the federal agency responsible for all census and statistical surveys in Brazil. It coordinates the National Statistical System, embracing data collection and dissemination for social, population and economics statistics. It also maintains and disseminates data about cartography, geography and natural resources in the country.

IBGE has about 8000 employs with headquarters in Rio de Janeiro and local offices in all the Brazilian States.

2. EDITING APPLICATION GENERATOR: CRIPTAX

2.1 Historical

Data collection for most of the surveys conducted by IBGE is still done primarily by interviewers or enumerators. Enumerators obtain the data using "paper and pencil interviewing" (PAPI).

Data entry at IBGE for census and complex surveys, has been performed in a "heads down" mode without error detection at time of entry. Due to questionnaire's complexity and data entry mode, the "data review" phase can be complex requiring many editing rules for the different levels of error detection.

Formerly, the editing rules were mapped into editing applications by a data processing specialist, using tailored programs. As the rules were specified without any standardization we had communication problems between the domain experts and the application developers.

After analyzing a number of packages, commercial and institutional ones, we decided to develop a specific editing generator called "CRIPTA". Its primary objectives were:

- a standard language definition for the editing rules specification;
- automatic application code generation, directly from the editing rules specified;
- a standard language definition for the editing program specification.

Subsequently, with the evolving technology allowing decentralization and portability, we developed a new version of CRIPTA, called CRIPTAX, which generates portable 'C' code applications and enables the integration with a set of generalized software packages used for the different phases of survey's processing cycle.

2.2 CRIPTAX Language

Is a pseudo code language, Portuguese like, composed by a set of procedural blocks and automatic functions which enables the specification of the editing rules and the editing program.

The procedural blocks and the automatic functions used at the editing program specification allows: to generate new variables by grouping or recording variables; to access variables from different kind of records; to calculate frequencies/totals and to incorporate, automatically, a set of editing rules.

2.3 CRIPTAX Generator

CRIPTAX generates an editing application from the three basic entities: Survey Dictionary, Survey Editing Rules and Survey Editing Program Specification. It runs under Mainframe environment and access the survey dictionary (file layout) and survey editing rules from IBGE's Meta-Database.

The system enables the generation of editing applications for different environments: PL/I is generated for mainframe environment, and C code for UNIX and Windows.

[1] Reina Marta Hanono, IBGE-Brazilian Institute of Geography and Statistics, Av. Chile, 500-12 andar - Rio de Janeiro - RJ - CEP: 20031-170 - Brazil; e-mail: rhanono@ibge.gov.br.

2.4 Editing Rules Specification

A standard specification for an editing rule is composed of a code error; a message error; a list of variables to be displayed; a condition to detect the error; and an effect (action to be performed).

The **condition** can be a combination of logical expressions, and CRIPTAX functions, connected by relational and logical operators. The **effect** is a set of CRIPTAX attribution commands, and the commands **ERR** (to display the associated variables in case of error) and **FORGET** (to abandon the record).

The CRIPTAX editing rules language allows subject matter specialists with little knowledge of computing to specify the editing checks to be applied to the data. In case an editing rule is required, which is beyond the capability of the automatic generation of CRIPTAX, then the rule is described in Portuguese, and the developer expert codes a subroutine which must be associated to it's code error in the Editing application specification.

2.5 Editing Application

Using the procedural blocks of CRIPTAX language, the user defines the way in which the questionnaire will be processed; by record or by questionnaire, and the titles and options for tailored report generation. It also commands the inclusion of the editing rules to be applied for each kind of record.

The associated subroutines to the editing rules, specified in natural language, must be placed in the editing program, using CRIPTAX commands or calling external procedures in another language.

The system generates two types of editing programs: a batch program, which produces a printed standard report; and an on-line program, which allows the correction task at time of editing cycling the questionnaire while errors are detected. In both cases the program displays the record identification, along with the detected errors, and the set of variables which can be modified in order to correct the error.

Another option of the system is the generation of an editing sub-routine, callable by the user's front-end application.

The generated editing application must be supplied by a data access routine (named in the dictionary) in case the survey data are stored in a DBMS. This enables the generated editing application to deal with survey data stored in different DBMS, and to run under a Client/Server architecture.

2.6 Process Improvements

CRIPTAX introduced a change in the way of working. It allows the subject matter specialists, to be directly responsible for the specification of the editing rules. In this way the user's knowledge about editing is incorporated early in the development phase, greatly reducing the communication problems between the subject matter and data processing staffs.

Now the subject matter specialists are responsible for the application correctness leaving the data processing specialists responsible for the application performance and feasibility.

2.7 Usage at IBGE

CRIPTAX has been used for the generation of the editing application for the Demographic Census 1991, the National Household Surveys ('PNAD') 1992-1996 and for the Agricultural and Economical Census 1995 in a UNIX environment.

The generated editing routines by CRIPTAX were used at the Annual Industrial Survey (PIA) 1992-1995 in mainframe environment accessing DB2 data with a CSP front-end application.

At present, some small surveys use the editing routine generated by CRIPTAX, called by a Delphi or a Visual Basic front-end application, in a Windows environment.

Specific training was provided for each development team in the specification of the Survey Dictionary and Editing Rules. The data processing specialists were trained in the specification of CRIPTAX programs and routines associated with editing rules specified in natural language. Each training module requires about fifteen hours.

In all implementations, we realized a significant reduction of time for the application development, reduction of specification errors, and improvement in user satisfaction.

3. CODING APPLICATION GENERATOR: SISCOD

3.1 Purpose

The basic goal of a computerized coding application is to assign codes to the various descriptions in a survey based on standard classifications.

In order to improve performance, the texts associated with each code must be analysed and reduced to a standard format to be used for comparison. This mapping of codes to texts is performed based on a set of specific parser rules which will be applied to both the standard classifications and to the text variables to be coded from the questionnaires.

3.2 Description

SISCOD is a Generalized Coding Application Generator, which enables the user to generate a coding application given a file containing the standard classification and a survey data file.

SISCOD works in two phases in order to generate a coding application: the Code Database generation and test; and the Coding application generation. These are described below.

3.2.1 Code Database Generation and Test

In this phase, the user informs the number of descriptions to be coded in the generated application, since it allows the coding of different descriptions, like religion, occupation, city *etc.* at the same application.

For each description to be coded the user must provide the sequential reference file containing the codes and associated texts and its layout and the parser rules.

The system generates a code database applying the, previously specified, set of parsing rules to the standard texts in the sequential reference file. SISCOD uses a default set of parsing rules, such as plural and gender elimination, and phonemes replacement.

Once the set of parsing rules is defined, the user commands the generation of the code database. Before creating the code database the system performs, for each description, an analysis of the texts in the sequential reference file and reports statistics about the frequency of each analysed word, giving the number of codes in which it appears and the number of times it appears on each code. Based on this report, the user can choose to change the specified set of parser rules, adding or eliminating words and separators or even changing the sequential reference file of the description.

If the set of parser rules or the sequential reference file is changed, the system performs the analysis again until the user commands it to generate the code database.

Once the code database has been generated, the user can run tests on it. For this task, the system allows him to enter a text, in the screen, and observe the system performance trying to code it. If the typed text has only one correspondent code in the database the system will show it. Otherwise it will show the multiple codes associated with the text or will inform that there is no match.

After performing the tests the user can either go back to the previous phase to regenerate the code database or proceed to the next phase to generate the coding application.

The advantage of this test phase is that the developer can test the code database independent of the survey data file.

3.2.2 Coding Application

Once the code data base is generated, the user is able to generate the Coding Application. This task has an easy menu driven facility.

The coding application is composed of two modules, the automatic and the assisted ones.

The Automatic Coding Application is batch oriented and resolves the mapping of text-to-code, for one or more descriptions at a time, tagging the unresolved records.

The Assisted Coding Application is an interactive module. It displays the unresolved records, allowing the user to select a code in case of multiple matching or to change the questionnaire text when there is no matching.

In both cases, Automated and Assisted Coding, the system allows the user to introduce a set of consistency rules, to be checked at time of coding. The consistency rules are formed by a set of editing rules specified in CRIPTAX.

In terms of achieving generality the survey database access routine must be specified by the user. Nevertheless, the system gives a standard schema for its specification.

3.2.3 Coding Application Production

The generated coding application can be used simultaneously by many users. The recommended way to use it is to perform the Automatic Coding Application followed by the Assisted Coding Application. If necessary, during the production phase the generated application can be adapted to the immediate needs of the users without software modification, by updating the Code Database.

It is worth emphasizing that both applications use the same parsing techniques (defined by the user) for a questionnaire text and search algorithms.

3.3 Coding Algorithm and Searching Technique

While performing the coding, after picking up the text from the questionnaire, both applications divide it into words, apply the parsing and phonetics rules on the words and create strings to be used for the searching. Then they perform the search in the word file with the most frequent words, created during the generation of the Code Database. In case a match isn't found, a search in the description file, created during the generation of the Code Database, is done. After the search a code is returned indicating whether it was coded or not. This code is used by the Assisted Coding Application, in order to access only the non-coded texts.

3.4 Usage at IBGE

SISCOD generated applications have been used with several surveys at IBGE. It was used for the Demographic Census 1991, for all the description to be coded, such as: religion, activity, occupation, migration, country of birth and instruction level. It was also used in the Annual National Household Survey "PNAD", in 92, 93 and 95 for occupation, activity and migration.

After analysing both cases we verified that the efficiency was of about 75% for all the descriptions, except occupation and activity (which are really more complex),and the accuracy (the percentage of well coded records) was satisfactory in terms of the sample analysed. It is worth to emphasize that the efficiency (the percentage of records automatically coded) depends on the Code Database input and the description subject.

3.5 The Software Package

SISCOD runs under UNIX environment. It requires "OPEN BASE", Data Base Management System, its associated language Opus and a 'C' compiler.

REFERENCES

Actr – Automated coding by text recognition (Sept/1990). Version 2. *System Overview*, Statistics Canada.

Barbosa, D.M.R., and Hanono, R.M. (Jan/Jun 1988). Estudo das ferramentas para apuração de dados. *Revista Brasileira de Estatística*, 85-100, Rio de Janeiro, 49(191).

Hanono, R.M., and Barbosa, D.M.R. (Sept/1992). A tool for the automatic generation of data editing and imputation application for survey processing. *Survey and Statistical Computing* (SGCSA) – North Holland, 449-456.

Silva, A.C.M., Hanono, R.M., and Barbosa, D.M.R. (July/1993). A Tool for the Automatic Generation of Computer-assisted Coding Application. Submitted and accepted by SGCSA - *Computerizing Survey Support Systems*.

Silva, P.L.N., and Bianchini, Z.M. (July/1993). Data Editing Issues and Strategies at the Brazilian Central Statistical Office.

SESSION C-2

Topics in Estimation and Variance

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

COMPARING ESTIMATION METHODS FOR A MONTHLY BUSINESS INQUIRY

P. Kokic[1] and T. Jones

ABSTRACT

Two different forms of estimators have been used in repeated monthly business surveys in the UK Office for National Statistics, one based on movements in matching units, which ignores unmatched sample members and projects forward from an initial base estimate. The second produces separate estimates each month using auxiliary information. A third compromise option also considered here is based on estimating the month on month change using both the matched units and auxiliary information. A method of benchmarking to annual estimates is also explored as a technique for improving the precision of both matched pairs estimators.

KEY WORDS: Ratio estimator; Matched pairs; Benchmarking; Permanent random number.

1. INTRODUCTION

In the UK Office for National Statistics (ONS) two different methods of estimation are used in monthly business surveys. One method is called matched pairs estimation, where the change between two consecutive months is estimated by the average relative change of units that are in the common sample for the two months. To produce an estimate of level (total) an initial base level estimate is used. The matched pairs approach has been used because it produces smooth estimates of month on month change, but there is a general suspicion, particularly due to the cumulative effects of register births, deaths and non-response, that longer term estimates of change and hence level estimates produced by this method are poor. Simulation results presented in this paper confirm this suspicion.

The second approach often used in ONS monthly surveys is ratio estimation (Cochran 1977). Ratio estimation is a standard method of producing estimates of level. However, particularly with data that is strongly correlated over time, estimates of month on month changes are likely to be more volatile than for the matched pairs approach described above.

One way of combining the strengths of both estimation methods above is to use a benchmarking procedure. Essentially, annual estimates of total are produced by each method then adjustments are made to ensure that the matched pairs annual total corresponds to the ratio estimate annual total.

In the following section we precisely define the various estimation methods used. In Section 3 a simulation study based on Retail Sales Inquiry data is described and results are presented in Section 4. Finally conclusions are drawn in Section 5.

2. ESTIMATION METHODOLOGY

Suppose at time point $t = 1, 2, K, T$ within stratum $h = 1, 2, K, H$ there are N_{th} units in a finite population. In each stratum h a simple random sample of size n_{th}, denoted by s_{th}, is selected and a variable y is observed for the sampled units only. Associated with this variable y is a single auxiliary variable x, which is assumed to be known for all units in the population at each time point.

The aim is to produce accurate estimates of both the total,

$$T_t = \sum_{h=1}^{H} \sum_{i=1}^{N_{ht}} Y_{thi},$$

and the relative change in the total between time points $r < t$,

$$I_{rt} = \frac{T_t - T_r}{T_r} \tag{1}$$

with a single estimation methodology. Practical interest, however, is concerned mainly with changes in the most recent 12 month period although longer term changes are also important.

2.1 The Various Estimators

The straightforward within-stratum ratio estimator is often used for estimating a finite population total, and when x and y are strongly correlated it is well know to produce quite efficient estimates of the population total. It is defined as

$$\hat{T}_{Rt} = \sum_{h=1}^{H} X_{th} \frac{y_{ts_{th}}}{x_{ts_{th}}},$$

where X_{th} is the sum of the x-values over the population units in h and $x_{ts_{th}}$ and $y_{ts_{th}}$ are the corresponding sums over the sampled x and y units, respectively.

One method of producing a more efficient estimate of the difference $T_t - T_{t-1}$, the numerator of (1), is to use a

[1] Philip Kokic, Department of Social Statistics, University of Southampton, Southampton, S017 1B7, U.K., andTim Jones, Office for National Statistics, U.K.

composite estimator (Cochran 1977, p. 346-355). In composite estimation two estimators of the difference are formed, one based on the sample in common between the time points t and $t-1$, which we will denote by s_{cth}, and one based on the remain (non-overlapping) sample. Both these estimates are weighted together according to the inverse of their respective variances and a more efficient estimator of change is produced.

In ONSs Monthly Retail Sales Inquiry (RSI) the amount of monthly sample rotation is small and the degree of correlation of responses between consecutive months is very high, often in excess of 0.95. In this case very efficient estimates of change can be formed by just using the sample in common from one time point to the next. Little additional benefit would be obtained by using the unmatched sample. This suggests that an estimator of the form

$$\hat{T}_{Mth} = \hat{T}_{M,t-1,h} \frac{y_{ts_{cth}}}{y_{t-1,s_{cth}}} \quad \text{where } \hat{T}_{Mt} = \sum_{h=1}^{H} \hat{T}_{Mth} \qquad (2)$$

would be highly efficient for differences. To be applied in practice some initial starting estimate must be supplied. One possibility is to use $\hat{T}_{D1} = \hat{T}_{R1}$. We shall refer to (2) as the matched pairs estimator with ratio updating. With this approach the changes are estimated accurately as the influence of extreme observation in the unmatched portion of the sample is avoided. However, the level estimate can drift over time, especially as the effects of births, deaths and changing auxiliary values are largely excluded.

One way of taking the auxiliary information into account and possibly produce more accurate estimates of level is to update the matched pairs estimator by estimating the total difference between consecutive months. That is, let

$$\hat{T}_{Dt} = \sum_{h=1}^{H} X_{t-1,h} \frac{d_{ts_{cth}}}{x_{t-1,s_{cth}}} + \hat{T}_{D,t-1},$$

where $d_{ts_{cth}}$ is the sum over the common sample s_{cth} of the response differences $d_{ti} = y_{ti} - y_{t-1,i}$.

2.2 Benchmarking

Both of the matched pairs procedures described above are specifically designed to estimate change accurately. However, because they may not produce good estimates of level in the longer term, regular adjustments should be performed. If there was some way of gradually adjusting these estimates so that over some given time frame they closely agreed with the ratio estimator of total, then the resulting estimator should have desirable properties for estimating both level and change. If this adjustment is performed on an annual basis, we shall refer to it as an annual benchmarking procedure.

The procedure that is used by ONS for benchmarking its estimates is based on splining methods (Baxter 1994). The method ensures that the smoothness of the initial series is maintained and it is relatively straightforward to apply in practice. Only this method of benchmarking will be considered although alternative techniques are available (Durbin and Quenneville 1997).

In the past, it has been usual practice to benchmark estimates with the results of an annual inquiry. However it is often necessary to wait for over two years from the beginning of the year to which the annual inquiry relates before the results are available. The alternative procedure used in this paper has the advantage that it could take place every 12 months as soon as the data for the twelfth month of the year in question has been collected. The disadvantage, as with any benchmarking procedure, is that revisions to estimates must be made as addition annual benchmarks become available.

3. SIMULATION STUDY

To assess the precision of the various estimators proposed in the previous section and to examine the properties of the benchmarked estimators a large scale simulation study was undertaken. Real data from the RSI was modelled and the simulation procedures were constructed in such a way as to reflect the sample rotation system as used in ONS.

3.1 Survey Data

The current sample design of the RSI is briefly as follows. The outcome variable is average weekly sales. The auxiliary variable is turnover. Its value is stored on the ONS businesses register. It is an annual total derived from value added tax returns and is closely linked to sales. The survey is stratified by 27 industry strata and 6 size strata according to register turnover. The top 1-2 size strata are completely enumerated and within stratum ratio estimation is used in the remaining strata. The sampling method is essentially fixed size simple random sampling, but the samples are selected using a permanent random number (PRN) rotation system very similar to that used by the Australian Bureau of Statistics (Hinde and Young 1985). RSI survey data from October 1995 until February 1996 were used in modelling the population. Over this period the average sample size was 4881 enterprises and the average population size was 226,550 enterprises.

3.2 Modelling the Population

The actual turnover auxiliary variable was available for all 226,550 population units and so this was used. This variable was held fixed over the whole simulation period. The sales data (y) within each month was modelled according to the lognormal regression model: $\ln(y) = \alpha + \beta . \ln(x) + \varepsilon$, where x is register turnover and ε are independent observations from a normal random variable with mean 0 and variance σ^2. Analysis of the RSI data suggested a very high correlation (ρ) of the residuals from the model above between consecutive months. Its value was estimated as slightly higher than 0.95 across all industries, however, for simplicity the value 0.95 was used instead. The parameters in the model were estimated separately within each broad industry category using robust estimation procedures.

3.3 The Simulation Procedure

Simulations covered a period of $T = 96$ months so 96 months of population data were constructed. The model was used to generate a set of values for all units in every month. This single set of values remained fixed throughout the simulation process. The procedure for simulation was as follows.

(a) Each unit was given a PRN at random and then sorted within strata according to the PRN.

(b) The sample for each month was selected using a rotation period of 15 months.

(c) The process above was repeated independently for each of 250 simulations. This produced 250 data sets of 4,881 observations for each of the 96 months.

From these data various estimates of level were produced for each of the monthly samples generated in each simulation. Benchmarking was performed on the two matched pairs estimates (at broad industry level). Both estimates of level and estimates of change relative to the previous months' estimates were computed. These were compared in terms of their root mean squared errors.

4. RESULTS

Figure 1 shows the root mean squared error (RMSE) of the non-benchmarked estimators of population total across the whole retail sector, while Figure 2 shows their RMSEs for estimating an n-month change when the reference base month in (1) is $r = 24$ and $t = n + r$. From Figure 1, one can clearly see the potential for disaster with either of the matched pairs (MP) estimators when estimating total. Both these estimators have a tendency to wander off course.

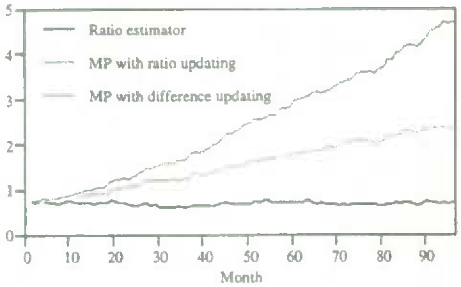


Figure 1. RMSE of level estimates for the whole retail sector. Percent of true total.

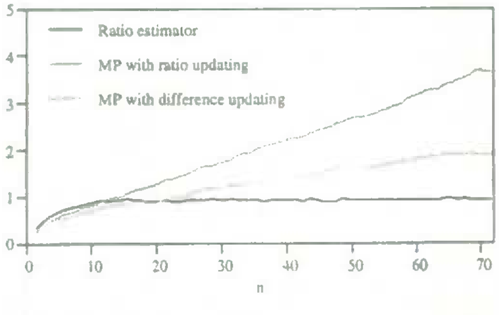


Figure 2. RMSE of n-month percentage change estimates for the whole retail sector. Reference base month is $r = 24$.

For estimating short-term change the MP estimators are more precise than the ratio estimator (Figure 2). However, both MP estimators grow progressively worse as the length of the comparison period (n) increases, whereas the precision of the ratio estimator levels off beyond $n = 15$ months.

Once benchmarking is performed the RMSE performance of both MP estimators for estimating level improves considerably (Figure 3). For the majority of the time period under consideration, the estimator with the smallest RMSE is the benchmarked MP estimator with difference updating. The corresponding estimator with ratio updating initially has about the same RMSE, but its performance deteriorates significantly over time. In the final 12 months no benchmarking can be performed and as can be seen the precision of both MP estimators declines rapidly over time until both are considerably worse than the standard ratio estimator.

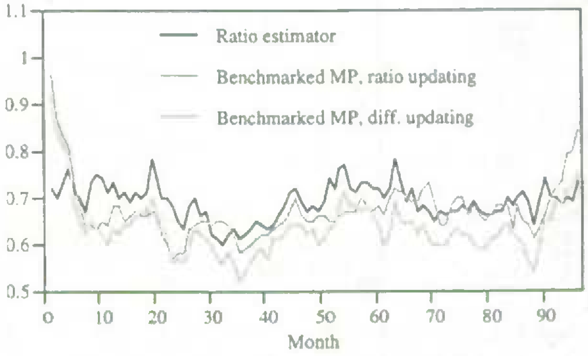


Figure 3. RMSE of level estimates for the whole retail sector. Percent of true total.

Figure 4 shows the RMSEs of estimates of month on month changes. Clearly the MP estimator with difference updating is more accurate for estimates of change than the ratio estimator. Also note the deterioration over time of the RMSE performance of the MP estimator with ratio updating. This deterioration in performance is most likely due to the lagged effect of the multiplicative updating term in (2). The reason why the effects of these updating terms are not being removed could be due to the fact that the benchmarking has been performed at a level higher than individual strata.

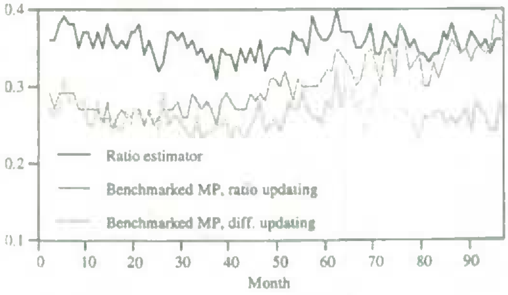


Figure 4. RMSE of 1-month percentage change estimates for the whole retail sector.

Figure 5 shows the RMSEs of the *n*-month change estimates when the reference base month is $r = 24$ (as in Figure 2). When *n* is 12 or less the RMSEs of both benchmarked matched pairs estimators is considerably less than the RMSE of the ratio estimator. The performance of all three estimators is about the same for n larger than 17.

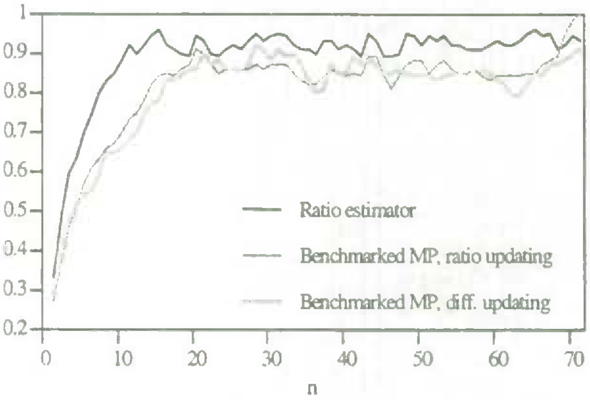


Figure 5. RMSE of n-month percentage change estimates for the whole retail sector. Reference base month is $r = 24$.

5. CONCLUSION

The results presented in this paper confirm the characteristics thought to be associated with the matched pairs methodology: Short term changes are estimated with greater precision, but there is a marked tendency for the estimates of levels (and hence longer-term changes) to wander off course.

The results suggest that this method of benchmarking the MP estimators leads to estimates of levels which are more accurate even than the simple ratio estimator, except possibly at the very end of the series. The cost of this improvement lies in the need to revise the estimates every year. A further cost is that it is likely to be difficult to estimate the variance of the new estimator.

The MP estimator that uses ratio updating appears to perform less well in the longer term than the alternative estimator based on differences. This probably reflects the level at which benchmarking was applied in the study. It is likely that benchmarking at the level of individual strata would eliminate this effect, but this remains to be tested.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*. Third edition, New York: John Wiley.

Durbin, J., and Quenneville, B. (1977). Benchmarking by state space models. *International Statistical Review*, 65, 23-48.

Hinde, R., and Young, D. (1985). Synchronised sampling and overlap control manual. Australian Bureau of Statistics internal document.

Hughes, P.J. (1988). Design and Estimation Issues for Rotating Business Surveys. Ph. D. thesis, Department of Social Statistics, University of Southampton.

Hughes, P.J. (1991). Design for Composite estimation with changing survey frames. *Journal of Official Statistics*, 7, 77-91.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

ESTIMATION OF VARIANCE IN PRESENCE OF IMPUTATION

F. Gagnon, H. Lee, M. Provost, E. Rancourt and C.-E. Särndal[1]

ABSTRACT

In sample surveys and censuses, imputation is commonly used to compensate for nonresponse. However, imputation introduces error in the point estimates, in addition to the sampling error. The fact that imputation may cause bias of the point estimate has traditionally been the main concern and the impact of imputation on the variance has largely been ignored, particularly for single imputation. As a result, the total variance is often wrongly estimated. It is now increasingly being recognized that the error due to imputation must also be taken into account, because it can be a significant portion of the total variance. This can be seen in the emerging literature in this area where a number of different approaches have been proposed. It is now possible to correctly estimate the total variance of Horvitz-Thompson, ratio and regression estimators when various imputation methods have been applied.

This paper reviews the model-assisted approach to the variance estimation problem for the Generalized Regression Estimator (GREG) in presence of single imputation. The method is currently being developed to be incorporated into Statistics Canada's Generalized Estimation System (GES). It also discusses how an assessment of the imputation variance can help improving survey quality.

KEY WORDS: Completed data set; Nearest neighbour imputation; SIMPVAR; Single imputation.

1. INTRODUCTION

The problem of estimation of the variance in presence of imputation is drawing more and more attention. As can be seen from the literature, there are now several variance estimation methods available to handle the problem. They are the model assisted method, (Särndal 1992), multiple imputation, (Rubin 1978, 1987), the two-phase approach, (Rao and Sitter 1995), the jackknife technique, (Rao and Shao 1992) and the bootstrap, (Shao and Sitter 1996). These methods are providing theoretical grounds for building estimation systems that estimate the variance in a manner that is more "correct" in the sense of also taking imputation into account. For instance, general requirements for incorporation of the model assisted method and the jackknife technique into the Generalized Estimation System are presented in Lee, Rancourt and Särndal (1995).

Parallel to this development, there has been an increased awareness among survey methodologists and survey practitioners about the possible impacts of imputation on estimation. Although it is recognized that the theory exists and could be useful, operational constraints are mentioned as the hindrance to the implementation of the available methods. As a result, there appears to be a gap between theoretical developments and survey practices.

Carrying out surveys is a complex task and no matter how much time and money is devoted to assuring a high quality, there will unfortunately be errors. Since estimates are produced from a process which by design is not error free, it is necessary to calculate the variance because it:

(i) Provides survey designers with a measure of the quality of the estimates;
(ii) Helps analysts to draw correct conclusions;
(iii) Enables statistical agencies to correctly inform users of data quality (*cf.* Policy on Informing Users on Data Quality and Methodology, (Statistics Canada 1992)).

When there is imputation, or any other survey operation that introduces error, it is particularly important to correctly estimate the variance so that the above three purposes are fulfilled. General practice has been to apply the ordinary variance estimator to the data set after imputation. But as will be seen in sections 3 and 4, this procedure is not sufficient since it incorrectly estimates the sampling variance and it completely misses the imputation variance. Survey methodologists are aware of these shortcomings of the ordinary method and would gladly consider using improved methods to obtain more correct estimate the variance. However, often cited reasons for not taking imputation into account when estimating the variance are:

(a) Correct methods are not readily available;
(b) Incorporation of the imputation variance would be useful but is not necessary;
(c) The procedure is too complex.

In this paper, we attempt to alleviate such misconception and bridge the gap between the theoretical developments and their practical application. We tackle the problem within the framework of the Generalized Estimation System (GES) which is used in many surveys at Statistics Canada.

[1] François Gagnon and Martin Provost, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; Hyunshik Lee, Eric Rancourt and Carl-Erik Särndal, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

We will confine ourselves to the situation where a completed data set is created using a single imputation method.

The paper is structured as follows: in Section 2, a short description of GES with its current developments is presented. A general framework for estimating the variance in presence of imputation is described in Section 3. Then, formulae for the case of nearest neighbour imputation are given in Section 4 and some examples are discussed. It is followed by concluding remarks in Section 5.

2. GES AND IMPUTATION

This section first presents a brief summary of the features of Statistics Canada's Generalized Estimation System (GES). Furthermore, the approach used to implement the method is outlined and SIMPVAR, a System for IMPutation VARiance to be appended to GES, is introduced.

2.1 Description of the Generalized Estimation System (GES)

The Generalized Estimation System, Estevao, Hidiroglou and Särndal (1995), is a microcomputer package that has been designed by Statistics Canada to produce domain estimates and corresponding variance estimates of population parameters such as totals, means and ratios. GES has been designed to accommodate several sampling designs. It is based on the model assisted theory of the generalized regression (GREG) estimator which uses auxiliary information to strengthen the quality of the estimates. This flexible framework permits the specification of a wide family of estimators including the traditional estimators such as Horvitz-Thompson, ratio, multiple regression, post-stratified, and many others. Two options are available for variance estimation: the Taylor linearization technique, for which GES is used the most, and the jackknife method.

GES is being used in many surveys conducted by Statistics Canada and, more recently, by other statistical agencies in the world. It has been developed for use with complete data files; however, missing data occur in almost all surveys, resulting in incomplete files. A common way to handle this problem is to impute data where nonresponse occurs. An important advantage of imputation is that it results in a complete data file that can then be used by GES to produce survey estimates. Nevertheless, imputation has some negative impacts that should not be ignored: potential bias and increased total variance. Potential bias that can be introduced by imputation may be largely reduced by using an appropriate imputation process. However, treating imputed data as collected data often leads to an understatement of the true variance, as described in detail in section 3. This underestimation might be more or less severe, depending on the imputation method used, the response rate for the variable, the domains defined by the user, as well as on some other survey parameters mentioned in section 4.

At present, no generalized estimation system, including GES, takes the impact of imputation into account when estimating variance. Knowing that the variance due to imputation may account for a significant portion of the total variance, an important goal is to incorporate the method described in section 3 into GES so as to permit correct estimation of the total variance.

2.2 Methods and Parameters Considered

Over the last few years, methods to take account of the variance due to imputation have been developed using the model assisted approach. Although we are interested in providing total variance estimates for all situations currently handled by GES, priority has been given to developing formulae for the cases most frequently encountered in the surveys conducted by Statistics Canada. To be more specific, methods to calculate the total variance for the GREG estimator in presence of imputation now exist for domain estimation of a total when ratio or nearest neighbour imputation has been used. These methods have been tested through simulations in the context of a stratified one-stage design where the imputation classes and the model groups for estimation were the same as the strata. The methods provide correct estimation of the total variance in presence of imputation for the estimation of a total for any domain. Extension of these methods will be developed in order to correctly estimate the variance in presence of imputation for all situations handled by GES.

2.3 Description of the System for IMPutation Variance (SIMPVAR)

The methods developed to take imputation into account in the variance estimation have been incorporated into a system called SIMPVAR (System for IMPutation VARiance). The preliminary version of this unofficial system is a first step into the process of building a module that will eventually be appended to GES to provide users with correct estimation of the total variance in presence of imputation.

SIMPVAR is a pull-down menu system that allows estimation of the imputation variance and consequently the estimation of the total variance which is the sum of the sampling variance and the imputation variance (see section 3 for more details on the decomposition of total variance). It also calculates the new coefficient of variation (CV) using the estimated total variance and it gives the proportion of the total variance due to sampling and imputation, respectively. Information about these proportions would help survey managers to better allocate resources in order to minimize the total variance.

An important feature of SIMPVAR is that it can use a large quantity of information from GES as inputs. Sampling design, type of estimator, auxiliary variables, strata information, population parameters to be estimated, domains definitions, design weights, g-weights and GES point estimates along with their associated sampling variance estimates are as many SIMPVAR inputs that come directly from GES. In addition to these, SIMPVAR requires other inputs related to imputation: the imputation method, the imputation classes, the auxiliary variable used for imputation, the respondent flags and the donor identifiers (for donor imputation methods only). The current beta version of SIMPVAR can estimate the imputation variance when nearest neighbour imputation was used.

SIMPVAR is still at an early stage of development and thus is not officially supported yet. In the future, this user-friendly system will be appended to GES to provide users with total variance estimates, new CV's based on total variance and the relative importance of sampling and imputation components in the total variance estimates.

3. GENERAL FRAMEWORK FOR ESTIMATION OF THE VARIANCE IN PRESENCE OF IMPUTATION

A sample s is drawn from the population $U = \{1, ..., k, ..., N\}$ by a given sampling design. The inclusion probability and the sampling weight of unit k are denoted by π_k and $a_k = 1/\pi_k$, respectively. Consider a domain of interest, $U_d \subseteq U$, for which we seek to estimate the total of the variable of interest y, $Y_d = \sum_{U_d} y_k$. For full response, GES estimates Y_d by

$$\hat{Y}_d = \sum_{s_d} a_k g_k y_k$$

where $s_d = s \cap U_d$ and g_k is the g-weight of unit k computed on the basis of a known auxiliary vector total, $\vec{X} = \sum_U \vec{x}_k$.

Because of nonresponse, Y_k is observed not for the full sample s but only for a response set r where $r \subseteq s$. The nonresponse set is denoted $o = s - r$. For a missing value $y_k, k \in o$, an imputed value, $\hat{y}_k$, is created. The completed data set is $\{y_{\bullet k} : k \in s\}$, where

$$y_{\bullet k} = y_k \text{ if } k \in r$$
$$\hat{y}_k \text{ if } k \in o.$$

GES uses the completed data set to compute the "imputed estimator" of Y_d,

$$\hat{Y}_{\bullet d} = \sum_{sd} a_k g_k y_{\cdot k}.$$

Its total error can be decomposed into

$$\hat{Y}_{\bullet d} - Y_d = (\hat{Y}_d - Y_d) + (\hat{Y}_{\bullet d} - \hat{Y}_d)$$

where $\hat{Y}_d - Y_d$ is the sampling error and

$$\hat{Y}_{\bullet d} - \hat{Y}_d = \sum_{o_d} a_k g_k (\hat{y}_k - y_k)$$

is the imputation error, where $o_d = s_d - r_d = o \cap U_d$ is the set of nonresponding units in the domain U_d.

The variance estimator for $\hat{Y}_{\bullet d}$ to be installed in SIMPVAR (GES) is composed as

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}}$$

where $\hat{V}_{\text{SAM}}$ is an estimate of the sampling variance of $\hat{Y}_d$, and $\hat{V}_{\text{IMP}}$ is an estimate (calculated from the completed data set) of the variance of the imputation error $\hat{Y}_{\bullet d} - \hat{Y}_d$. The term $\hat{V}_{\text{MIX}}$ corresponds to the covariance between the sampling error and the imputation error. Although $\hat{V}_{\text{MIX}}$ also contributes to the total variance, we have found its contribution to be small in most cases. In SIMPVAR, $\hat{V}_{\text{MIX}}$ is used but it is simply incorporated into the imputation variance.

4. VARIANCE ESTIMATORS WHICH TAKE IMPUTATION INTO ACCOUNT

The calculation of $\hat{V}_{\text{SAM}}$ can be carried out by using the already-programmed variance estimator in GES, only it is calculated not on a fully observed sample of y-values, but on the completed data set. For some imputation methods however, transformations to the data may have to be performed for variance calculations as in Gagnon, Lee, Rancourt and Särndal (1996). This leads most of the time to a conservative estimation of the true sampling variance $\hat{V}_{\text{SAM}}$, but often the overestimation is not of significant proportion.

The calculation of $\hat{V}_{\text{IMP}}$ is based on model assisted reasoning. Assuming that the imputation is carried out using an auxiliary variable (which is not necessarily the same as that used for estimation), the imputation model, denoted ξ, is of the form

$$y_k = \beta z_k + \varepsilon_k,$$

with

$$E_\xi(\varepsilon_k) = 0;\ E_\xi(\varepsilon_k^2) = \sigma^2 z_k;\ \text{and}\ E_\xi(\varepsilon_k \varepsilon_{k'}) = 0 \text{ for } k \neq k'$$

where z_k is the auxiliary variable used to perform imputation. The $\hat{V}_{\text{IMP}}$ and $\hat{V}_{\text{MIX}}$ have to satisfy

$$E_\xi(\hat{V}_{\text{IMP}} - \hat{V}_{\text{IMP}}) = 0;\ \text{and}\ E_\xi(\hat{V}_{\text{MIX}} - \hat{V}_{\text{MIX}}) = 0.$$

For nearest neighbour imputation, GES provides a correct estimate of the sampling variance under most conditions. However, the imputation variance must be estimated using an extension of the estimators presented in Gagnon, Lee, Rancourt and Särndal (1996) and Rancourt, Särndal and Lee (1994):

$$\hat{V}_{\text{IMP}} = \left\{\sum_{l \in r} S_{ld}^2 z_l + \sum_{k \in o_d} w_k^2 z_k\right\}\hat{\sigma}^2$$

and

$$\hat{V}_{\text{MIX}} = \left\{\sum_{r_d} (w_k - 1)\ S_{kd} z_k + \sum_{o_d} w_k (w_k - 1) z_k\right\}\hat{\sigma}^2,$$

where S_{ld} is the sum of the weights for recipients k in domain d having l as their donor (Several recipients may have the same donor) and

$$\hat{\sigma}^2 = \sum_r \left(y_k - \frac{\bar{y}_r}{x_r} z_k \right) \Big/ \sum_r z_k .$$

Then, the variance estimator that we propose for nearest neighbour imputation is

$$\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}},$$

where $\hat{V}_{\text{IMP}}$ and $\hat{V}_{\text{MIX}}$ are given above and $\hat{V}_{\text{SAM}}$ is computed with the existing formula from the completed data set.

To assess the performance and illustrate the importance of the variance estimators that take imputation into account, simulation studies were carried out. For these, several artificial populations of size 2000 were generated. In 1,000 iterations, stratified simple random samples of size 400 were drawn from the population with sampling fractions varying between 10% and 100% depending upon the stratum. Nonresponse was generated using Bernoulli trials, which corresponds to a uniform response mechanism. The response rate was set at 50%. The missing data were imputed using nearest neighbour imputation and the estimation of the total was done using the ratio estimator Selected results for the case where the domain is the whole population are shown in Table 1.

Table 1

	Magnitude in (000,000)	Rel. Bias	% of $\hat{V}_{\text{TOT}}$
$\hat{V}_{\text{TOT}}$	28,03	—	—
$\hat{V}_{\text{SAM}}$	9,33	—	—
$\hat{V}_{\text{SAM}}$	9,27	-67%	33%
$\hat{V}_{\text{IMP}}$	19,00	—	68%
$\hat{V}_{\text{MIX}}$	-0.24	—	-1%
$\hat{V}_{\text{TOT}}$	27,79	-0.86%	—

From this example, it is clear that not taking imputation into account, that is, using only $\hat{V}_{\text{SAM}}$ to estimate $\hat{V}_{\text{TOT}}$ renders the estimate of the variance totally inadequate. However, Table 1 shows that the formula for $\hat{V}_{\text{IMP}}$ leads to an almost correct estimation of the total variance. It is also very useful to know the percentage of the imputation variance, as this could be used to better allocate the resources between an improved follow-up (which would reduce imputation) and a larger sample size (which would reduce sampling error).

5. CONCLUSION

In this paper, we have described the problem of variance estimation in presence of imputation. We have pointed out that despite the increasing awareness and needs for such methods, and the fact that there are methods available, there is still a gap between the theory and the methods in place for production. We have also pointed out that this gap is now closing, and that some procedures are now being incorporated into systems (SIMPVAR / GES).

Going back to the three reasons expressed for not taking imputation into account when estimating the variance (not necessary, not available, too complex) presented in Section 1, we would argue that they are no longer valid. Section 4 demonstrates, not only the usefulness, but also the necessity of taking imputation into account in estimation of the variance. Even in a time and budget constrained environment, one must not solely rely on "ordinary" methods, because they could mislead users into drawing wrong conclusions. This may necessitate redesign or new programs, which will no doubt be more costly and time consuming than correctly estimating the variance in the first place.

In addition to the three purposes of variance estimation (i), (ii), (iii), discussed in Section 1, others come into play when there is imputation. Then, taking the imputation variance into account contributes to

(iv) improved knowledge of the impact of imputation;
(v) providing a correct variance estimation;
(vi) better allocation of resources between sampling and edit and imputation (*cf.* Section 4).

The reason why methods were not readily available was perhaps legitimate up until recently, but cannot be said to be valid anymore. For instance, Section 2.3 describes how the model assisted method is implemented into the system called SIMPVAR.

Finally, the procedures are not complex and the formulae have been successfully coded into programs. Further, once coded, the method is simple to use.

6. ACKNOWLEDGMENTS

We would like to thank Greg MacLean for his invaluable help in the development of SIMPVAR.

REFERENCES

Estevao, V., Hidiroglou, M.A., and Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

Gagnon, F., Lee, H., Rancourt, E., and Särndal, C.-E. (1996). Estimating the variance of the generalized regression estimator in the presence of imputation for the generalized estimation system. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 151-156.

Lee, H., Rancourt, E., and Särndal,C.-E. (1995). Variance estimation in the presence of imputed data for the Generalized Estimation System. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 384-389.

Rancourt, E., Särndal, C.-E., and Lee, H. (1994). Estimation of the variance in presence of nearest neighbour imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 888-893.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rubin, D.B. (1978). Multiple imputation in sample surveys – a phenomenological Bayesian approach to nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 20-34.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

Statistics Canada (1992). Policy on Informing Users on Data Quality and Methodology.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

HIERARCHICAL BAYES SMALL AREA ESTIMATION USING MULTI-LEVEL MODELS

Y. You and J.N.K. Rao[1]

ABSTRACT

Standard multi-level models with random regression parameters are considered for small area estimation. We also extend the models by allowing unequal error variances or by assuming random effect models for both regression parameters and error variances. We present these models in a hierarchical Bayes framework and estimate a small area mean by its posterior mean. Posterior variance of the small area mean is used as a measure of precision of the estimate. It automatically takes into account the extra uncertainty associated with the hyperparameters in the multi-level model. Gibbs sampling is used to compute the posterior means and posterior variances of small area means. The procedure is illustrated through a simple example.

KEY WORDS: Gibbs sampling; Hierarchical Bayes; Multi-level model; Small area.

1. INTRODUCTION

Battese *et al.* (1988) proposed and applied a nested error regression model to provide small area estimates. The model takes the form:

$$Y_i = X_i\beta + v_{0i}1_{n_i} + e_i \qquad i = 1, ..., m \tag{1.1}$$

where Y_i is the vector of length n_i for the character of interest for the sampled units in the i-th small area, X_i is the $n_i \times p$ matrix of explanatory variables, $1_{n_i} = (1, ..., 1)^T$ is the unit vector of length n_i, β is a set of p fixed regression parameters, v_{0i} is a scalar random effect for the i-th small area with $E(v_{0i}) = 0$ and $V(v_{0i}) = \sigma_v^2$. The e_i's are assumed to be independent random error vectors with $E(e_i) = 0$ and $v(e_i) = \sigma_e^2 I_{n_i}$, where I_{n_i} is the $n_i \times n_i$ identity matrix. v_{0i} and e_i are also assumed independent. For the whole population the model (1.1) applies with n_i replaced by N_i the small area population size.

Holt and Moura (1993) extended the model (1.1) to a multi-level model by introducing random regression coefficients and explanatory variables at the small area level helping to explain differences between small areas. For $i = 1, ..., m$, the model can be stated as follows:

$$Y_i = X_i\beta_i + e_i, \; \beta_i = Z_i\gamma + v_i, \tag{1.2}$$

where β_i is the i-th small area random regression coefficients, Z_i is a $p \times q$ design matrix of small area level variables, γ is a vector of length q of fixed coefficients, and $v_i = (v_{i1}, ..., v_{ip})^T$ is the vector of length p of random effects for the i-th small area. The v_i's are independent and have a joint distribution within each small area with $E(v_i) = 0$ and $V(v_i) = \Phi$, where Φ is an unknown variance covariance matrix.

In this paper, we present the multi-level model in a hierarchical Bayes framework and extend the model to a more general multi-level model which allows unequal error variances and random effects for both small area regression parameters and sampling error variances. The small area mean μ_i is estimated by its posterior mean and its precision is measured by its posterior variance. Posterior variance automatically takes into account the extra uncertainty associated with the variance components and hyperparameters in the multi-level model. We use the Gibbs sampling method to compute the hierarchical Bayes estimates and the associated posterior variances. Section 2 presents the hierarchical Bayes multi-level models with different error variance models. Section 3 presents a small data analysis. And section 4 is a conclusion.

2. HIERARCHICAL BAYES MULTI-LEVEL MODELS

2.1 Equal Error Variance Model

Suppose there are m small areas. Let y_{ij} denote the observation of a character of interest for the j-th unit of the i-th small area $(j = 1, ..., n_i)$. A hierarchical Bayes representation of model (1.2) is given by:

Model 1:

(i) $y_{ij} \mid \beta_i, \sigma_e^2 \underset{\sim}{\text{ind}} N(x_{ij}^T\beta_i, \sigma_e^2), (i = 1, ..., m; j = 1, ..., n_i)$;

(ii) $\beta_i \mid \gamma, \Phi \underset{\sim}{\text{ind}} N_p(Z_i\gamma, \Phi), (i = 1, ..., m)$;

(iii) Priors: $\gamma \sim N_q(0, D), \tau_e \sim G(\alpha, b)$ and $\Omega \sim W_p(\alpha, R)$ where $\tau_e = \sigma_e^{-2}, \Omega = \Phi^{-1}$, $G(a, b)$ is a gamma distribution with density $f(x) \propto x^{a-1}e^{-xb}, a > 0, b > 0, x \geq 0$, and $W_p(\alpha, R)$ is a Wishart distribution with density

[1] Yong You, Survey and Analysis Methods Development Section, HSMD, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

$$f(X) \propto |X|^{\frac{\alpha-p-1}{2}} \exp\{-\frac{1}{2}tr(RX)\}, X>0,\ R>0.$$

Remark 1.1: The prior distributions in (iii) are conjugate with the sampling and population distributions given by (i) and (ii) in the sense that they lead to full conditional distributions for γ, τ_e and Ω are again normal, gamma and Wishart distribution, respectively.

Remark 1.2: It is important to note that we have used proper priors on all the unknown parameters to ensure that all the posterior distributions are proper (Hobert and Casella 1996). Hence we do not face any problem of some posteriors being improper. Parameter values in the priors are chosen to reflect a fairly vague knowledge of the prior distributions.

Remark 1.3: In Model 1, we assume equal error variance σ_e^2 for all small areas. In practice, however, sampling error variances could be different for different small areas. A more general model should allow possibly different error variances.

2.2 Unequal Error Variance Model

In practice, it is more realistic to assume unequal error variances for the sampling errors. Let σ_i^2 be the true sampling error variance for the i-th small area. A straightforward extension of Model 1 leads to the following hierarchical Bayes multi-level unequal error variance model:

Model 2:

(i) $y_{ij}|\beta_i, \sigma_i^2 \ \underset{\sim}{\text{ind}}\ N(x_{ij}^T\beta_i, \sigma_i^2)$, $(i = 1, ..., m; j = 1, ..., n_i)$;

(ii) $\beta_i|\gamma, \Phi \ \underset{\sim}{\text{ind}}\ N_p(Z_i\gamma, \Phi)$, $(i = 1, ..., m)$;

(iii) Priors: $\gamma \sim N_q(0, D)$, τ_i iid $G(a, b)$, and $\Omega \sim W_p(\alpha, R)$, where $\tau_i = \sigma_i^{-2}$, $\Omega = \Phi^{-1}$.

Remark 2.1: Model 2 reduces to Model 1 when $\sigma_i^2 = \sigma_e^2$ for all i. From a hierarchical Bayes perspective, extension from the equal error variance model to the unequal error variance model is straightforward and also there is no difficulty in the Gibbs sampling implementation.

Remark 2.2: τ_i's are assumed to be independent and have the same prior distribution $G(a, b)$, where a and b are hyperparameters usually chosen to be very small to reflect a vague knowledge about τ_i.

2.3 Random Error Variance Model

In Model 2, we assume unequal error variances for the sampling errors. Kleffe and Rao (1992) used a random error variance model to derive the best linear unbiased predictor for small area means. A Bayesian extension of Model 2 due to Kleffe and Rao (1992) leads to the following Model 3, which allows us to consider random effects for both regression coefficient β_i and sampling error variance σ_i^2 for the corresponding i-th small area.

Model 3:

(i) Same as in Model 2;

(ii) Same as in Model 2;

(iii) $\tau_i|\eta, \lambda \ \underset{\sim}{\text{iid}}\ G(\eta, \lambda)$, $i = 1, ..., m$, where $\tau_i = \sigma_i^{-2}$;

(iv) Priors: $\tau \sim N_q(0, D)$, $\Omega \sim W_p(\alpha, R)$, $\eta \sim U^+$ and $\lambda \sim U^+$, where U^+ denotes a uniform distribution over a subset of R^+ with large but finite length.

Remark 3.1: In Model 3, we assume that τ_i's are iid gamma random variables with unknown parameters η and λ. Thus we have population models for both regression coefficient β_i and sampling variance σ_i^2. In Model 1 and Model 2, we only consider modelling β_i and put vague prior distributions on σ_e^2 or σ_i^2.

Remark 3.2: It is not easy to model the variance components τ_i. $G(\eta, \lambda)$ may not be a good population model for all τ_i's. Alternatively we can model τ_i in a more general way as β_i by specifying a regression model for the logarithm of τ_i. This may require some auxiliary information for the variance components. In the data analysis, we simply use $G(\eta, \lambda)$ as the population model for τ_i.

2.4 Bayesian Inference

For the three models, we are interested in finding the posterior distributions of β_i's given the data $Y = (\{y_{ij}\}, i = 1, ..., m; j = 1, ..., n_i)$, and in particular in finding the posterior estimation of small area means $\mu_i = \bar{X}_i^T\beta_i$, which depend on the estimation of β_i. Direct evaluation of the joint posterior distribution involves high-dimensional numerical integration, and is not computationally feasible. Thus we use the Gibbs sampling method (Gelfand and Smith 1990) to generate samples from the joint posterior distributions. Posterior mean and posterior variance of small area mean $\mu_i = \bar{X}_i^T\beta_i$ are estimated through the estimation of β_i by using the samples generated from the Gibbs sampler.

We fit the three models via Gibbs sampling using the BUGS program (Spiegelhalter, *et al.* 1995), aided by CODA Splus function (Best, *et al.* 1995) for assessing convergence and computing posterior summaries. The advantage of BUGS program is that it frees us from computational details.

3. DATA ANALYSIS

Following Holt and Moura (1993), we consider the small area estimation of household income in a county of Brazil. Our data set contains 10 small areas with 28 observations in each area. Let y_{ij} denote the j-th household's income in the i-th small area. The sampling model is given by

$$y_{ij} = x_{ij}^T\beta_i + e_{ij} = \beta_{0i} + x_{1ij}\beta_{1i} + x_{2ij}\beta_{2i} + e_{ij} \qquad (3.1)$$

where x_{1ij} denotes the number of rooms in the j-th household of small area i and x_{2ij} denotes the corresponding educational attainment of Head of Household. e_{ij} is the sampling error variable and its distribution is determined by the three error variance models discussed in Section 2. In model (3.1), β_i is the random regression coefficient corresponding to the i-th small area and is modelled as

$$\beta_{0i} = \gamma_0 + v_{0i}, \beta_{1i} = \gamma_{10} + \gamma_{11}z_i + v_{1i}, \beta_{2i} = \gamma_{20} + \gamma_{21}z_i + v_{2i}, \tag{3.2}$$

where $\gamma = (\gamma_0, \gamma_{10}, \gamma_{11}, \gamma_{20}, \gamma_{21})^T$ is the unknown regression parameters, $v_i = (v_{0i}, v_{1i}, v_{2i})^T$ is the i-th small area random effect vector distributed as $v_i \sim N_3(0, \Phi)$, and z_i is an area level variable defined as the average number of cars per household in each small area. Value of z_i is also centered around its overall sample mean.

To predict the small area mean $\mu_i = \bar{X}_i^T = \beta_{0i} + \bar{X}_{1i}\beta_{1i} + \bar{X}_{2i}\beta_{2i}$, where $\bar{X}_{1i}$ and $\bar{X}_{2i}$ are the i-th small area population means of the auxiliary variables x_1 and x_2, respectively, we first obtained the posterior estimate $\hat{\beta}_i$ for β_i, then μ_i is estimated as $\hat{\mu}_i = \bar{X}_i^T \hat{\beta}_i = \hat{\beta}_{0i} + \bar{X}_{1i}\hat{\beta}_{1i} + \bar{X}_{2i}\hat{\beta}_{2i}$. The standard error of $\hat{\mu}_i$ is defined as the squared root of the estimated posterior variance of μ_i, *i.e.*, $\hat{V}(\mu_i)$, which is given by $\hat{V}(\mu_i) = \bar{X}_i^T \cdot \hat{V}(\beta_i) \cdot \bar{X}_i$, where $\hat{V}(\beta_i)$ is the sample variance covariance matrix of β_i. We implemented the Gibbs sampler using the BUGS program. After a "burn-in" period of 2000 iterations, 5000 iterations were kept for analysis. Table 3.1 presents the estimated posterior small area means $\hat{\mu}_i$ with their standard errors. For comparison, sample mean $\bar{y}_i$'s are also included in the table.

From the results in Table 3.1 we noted that under Model 2 the estimated posterior means $\hat{\mu}_i$ have the smallest standard errors among the three models for all areas except area 5 and 8. The sample mean and estimated posterior area mean for area 5 are substantially larger than those values for the other areas, which indicates that area 5 is significantly different from the other areas and could be considered as an outlier area. The result of Model 3 is close to that of Model 1. For a better understanding of these results, it is necessary to look at the estimated error variances, since our models are based on different assumptions on error variances. Table 3.2 shows for each small area the sample variance and the estimated posterior error variances under the three models. For comparison, we also calculated the ordinary least square (OLS) estimates of the sampling error variances.

First we note from the OLS estimates that there are large variations among the ten areas. Model 1 assumes an equal error variance σ_e^2 for all areas. σ_e^2 is estimated by $\hat{\sigma}_e^2 = 76.76$, which is much smaller than the sample variances for many areas. Model 2 assumes unequal error variances σ_i^2 for all small areas. Under Model 2, the estimated error variances $\hat{\sigma}_i^2$ to some extent show the features of the areas. The most notable result is $\hat{\sigma}_8^2 = 160.09$, which indicates that there are large variations within small area 8. Model 3 assumes σ_i^2's are random variables with the same hyper population $G(\eta, \lambda)$. Under Model 3, all $\hat{\sigma}_i^2$ have moved toward $\hat{\sigma}_e^2 = 76.76$. This is the reason that in Table 3.1 the estimated posterior small area means under Model 3 are similar to those under Model 1. The result of Model 3 is between the results of Model 1 and Model 2. Model 2 assumes unequal error variances without any restrictions on σ_i^2's. Model 3 puts a model, a gamma distribution $G(\eta, \lambda)$, on σ_i^2's. The results in Table 3.1 and Table 3.2 show that $G(\eta, \lambda)$ is not a good model for σ_i^2's for the data set used here.

Table 3.1

Estimation of small area means

Small Area	Sample Mean	Estimated Posterior Small Area Mean (SE)					
		Model 1		Model 2		Model 3	
1	11.08	10.30	(0.98)	10.23	(0.82)	10.32	(0.90)
2	7.90	10.36	(1.06)	9.84	(0.86)	10.31	(1.01)
3	13.48	13.08	(1.09)	13.00	(1.09)	13.10	(1.07)
4	6.53	11.43	(1.26)	10.95	(1.10)	11.35	(1.19)
5	19.51	18.36	(1.29)	17.87	(1.58)	18.28	(1.33)
6	11.21	10.47	(0.96)	10.21	(0.93)	10.45	(0.95)
7	8.72	9.80	(0.99)	9.59	(0.98)	9.80	(0.97)
8	12.80	11.09	(1.20)	10.30	(1.21)	10.95	(1.21)
9	10.18	11.71	(1.11)	11.34	(1.00)	11.68	(1.08)
10	10.00	9.95	(0.96)	9.80	(0.87)	9.93	(0.90)

Table 3.2

Estimated small area sampling error variances

Small Area	OLS Estimates	Estimated Posterior Error Variances (SE)					
		Model 1		Model 2		Model 3	
1	38.17	76.76	(6.78)	40.10	(11.77)	66.68	(11.54)
2	31.75	76.76	(6.78)	34.19	(10.27)	65.86	(11.80)
3	81.26	76.76	(6.78)	94.50	(28.04)	78.16	(12.30)
4	48.73	76.76	(6.78)	52.08	(15.61)	69.58	(11.26)
5	115.98	76.76	(6.78)	121.71	(36.00)	84.35	(15.31)
6	90.74	76.76	(6.78)	94.04	(27.29)	78.35	(12.66)
7	101.67	76.76	(6.78)	102.37	(29.95)	80.20	(13.42)
8	135.65	76.76	(6.78)	160.09	(49.46)	90.94	(17.84)
9	59.10	76.76	(6.78)	63.46	(19.00)	71.72	(11.31)
10	62.86	76.76	(6.78)	65.88	(20.32)	72.25	(11.41)

In order to know how the data support each model, we calculated the so called conditional predictive ordinate (CPO) based on cross-validation predictive densities for each data point y_{ij} (Gelfand 1995). We present a CPO plot for the three models in Figure 3.1. Since CPOs are nothing but the observed likelihoods, larger CPOs suggest a more likely model. Clearly Model 2 is the best model, and Model 3 is slightly better than Model 1. Also there are many small CPO values for all three models, which indicates that the data set is far from our model assumptions. More work need to be done within each small area, especially for area 5 and 8.

5. CONCLUSION

In this paper we have presented hierarchical Bayes models for small area estimation using multi-level models. Clearly it is not easy to provide a suitable model for all small areas with satisfactory results, even the Markov chain Monte Carlo Bayesian methods such as the Gibbs sampling enable us to fit the data with Bayesian models of virtually unlimited complexity. The size and homogeneity of the areas and the availability of auxiliary information will affect

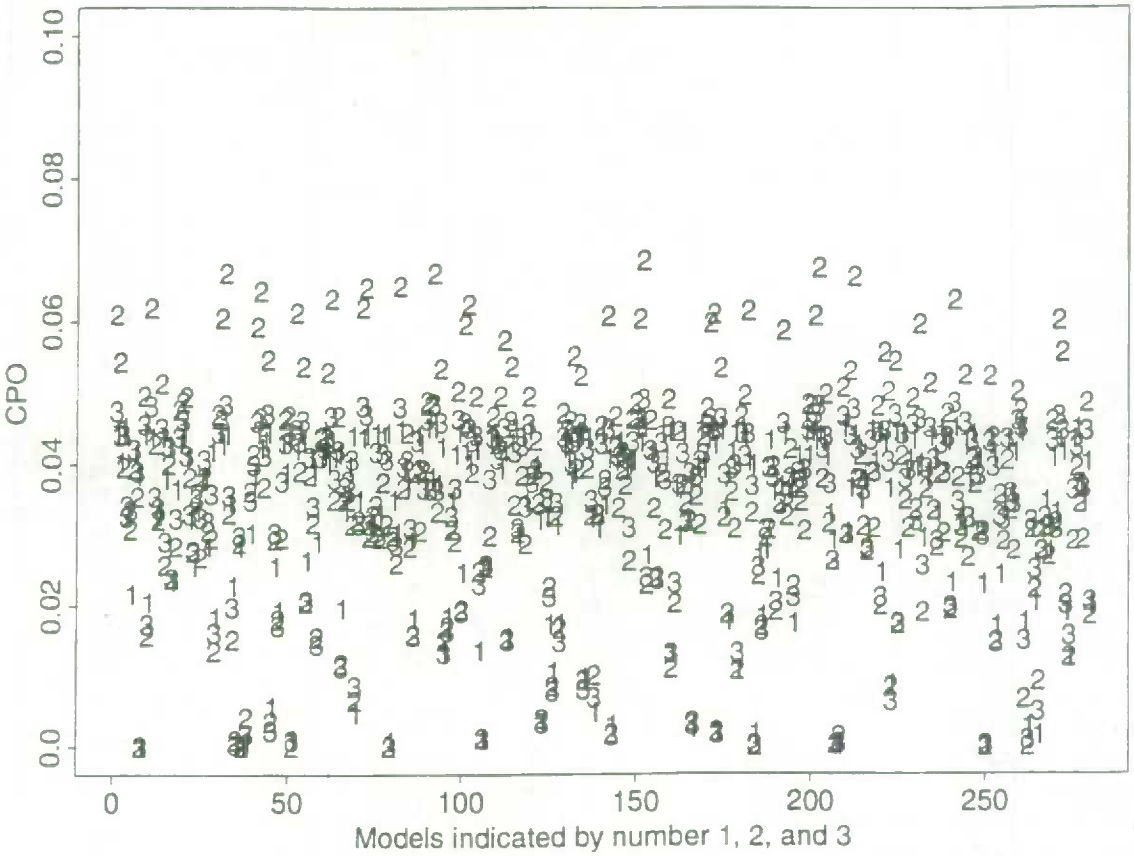


Figure 3.1. CPO Comparisons Plot

the final result. Models which prove suitable in some situations may be unsuitable in others. Nevertheless, the general hierarchical Bayes methodology is applicable to a wide variety of situations for estimation of small area parameters. Future work will look at the Bayesian model evaluation including the robustness of Bayes estimates and Bayesian model choice in more detail. It is also important to compare the hierarchical Bayes method with other methods used in small area estimation such as the empirical Bayes method and the empirical best linear unbiased predictor (EBLUP) method.

REFERENCES

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Best, N., Cowles, M.K., and Vines, K. (1996). CODA, Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30. *MRC Biostatistics Unit*, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

Dick, P., and You, Y. (1997). Bayes and census undercoverages. *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meetings*, Fredericton, New Brunswick, June, 1997.

Gelfand, A.E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice*. (Eds W.R. Gilks, S. Richardson, and D.J. Spiegelhalter), 145-161. London: Chapman & Hall.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.

Hobert, J.P. and Cassella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91, 1461-1473.

Holt, D., and Moura, F. (1993). Small area estimation using multi-level models. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1, 21-30.

Kleffe, J., and Rao, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under a random error variance linear model. *Journal of Multivariate Analysis*, 43, 1-15.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1996). *BUGS 0.5, Bayesian inference using Gibbs sampling manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.

SESSION C-3

International Experiences in Sample Design

TWO-PHASE SURVEYS OF ELUSIVE POPULATIONS

L. Barabesi[1], L. Fattorini and G. Ridolfi

ABSTRACT

A very common situation in environmental studies is determined when the population to be investigated constitutes a collection of objects on a study area. Usually, the inference problem is to estimate the total of an interest variable associated with each object (for example, objects may be shrubs or pollution sites, while the interest variable may be the berry quantity or the pollutant emission respectively). Aim of this paper is to analyze a two-stage estimator of the population total when: (i) objects are elusive so that the sampled units are those selected according to the well-known line transect design, (ii) the study area is too large to be surveyed entirely, so that only a sample of plots is considered.

KEY WORDS: Two-stage sampling; Line transect sampling; Population total estimation.

1. SAMPLING ELUSIVE POPULATIONS

Consider a collection of N elusive units on a delineated study area of size A. This usually happens when dealing with animal and plant populations. Denote by $x_1, x_2, ..., x_N$ the values associated with each unit and let the total X be the target parameter. Finally, denote by S the sample of distinct units selected from the population according to a line transect design, *i.e.*, those spotted along a transect line randomly thrown on the study area. Referring to units by their indexes, S obviously reduces to a subset of the population indexes $\{1, 2, ..., N\}$.

Quoting from Thompson (1992), suppose a rectangular study area with horizontal side (baseline) of length w and vertical side of length L. Then, the line transect design involves the following steps: (i) select a point at random on $(0,w)$ and throw a line of length L crossing the study area; (ii) include in S all the units spotted by the line. In this framework, the Horvitz-Thompson estimator of the population total

$$\hat{X} = \sum_{j\in S} \frac{X_j}{\pi_j} \tag{1}$$

can be performed, providing that the detection (selection) probabilities, say π_j $(j = 1, 2, ..., N)$, be positive for each unit and be readily quantified. Thus, in order to achieve estimates of type (1) the problem reduces itself to evaluate the π_j's.

If the Hayne (1949) detection model is supposed, in which the j-th unit is spotted when the observer enters the circle of radius r_j centered at the unit location, the detection probabilities turn out to be $\pi_j = 2r_j/w$, where r_j can be quantified straightforwardly by the radial distance from the observer to the detected unit. Obviously, if the j-th object is near the edges, a portion of the sightability circle may overlap the study area, so that the selection probability turns out to be $\pi_j = (r_j + c_j)/w$ where c_j represents the distance of the projection of the j-th object from the nearest edge of the baseline. Alternatively, the detection probabilities can be nicely quantified if the following assumptions are made on the process of sighting the units: (a) the probability of spotting any object on the area depends only on the perpendicular distance of the object from the line, (b) no object farther than a known distance $b < w/2$ is detected; (c) objects on the transect are spotted almost certainly. Now, suppose that the design is modified in such a way that whenever a transect is selected at a distance less than b from an edge of the baseline, an additional transect runs at the same distance from the other edge, outside the study area. Then, under the above mentioned assumptions, Thompson (1992) shows that all the objects have the same detection probability $\pi_0 = 2/\{wf(0)\}$, where f denotes the probability density of the distances of the detected objects from the line. If $f(0)$ is estimated from the sample by $\hat{f}(0)$, the resulting estimator obtained from (1) is not an Horvitz-Thompson estimator since the estimated values of the detection probabilities are used instead of the true ones. However it is straightforward to prove (see Fattorini 1995) that (1) is an unbiased estimator of X provided that $\hat{f}(0)$ is an unbiased estimator for $f(0)$, conditional on the sampled objects.

The problem of estimating $f(0)$ by the observed distances has been the crucial point of the statistical literature on line transect sampling. All the works in literature assume sample distances to be *iid*, while they only have the same marginal distributions (see Thompson 1992, p.194). However, Buckland *et al.* (1993, p.36) emphasize that usually the lack of independence among observed distances does not affect the bias of the $f(0)$ estimators, but is likely to have a heavy impact on their variances and, subsequently, on the sampling variance of (1). Accordingly, analytical expressions for this variance are unknown in this case.

Anyway, the problem of estimating the sampling variance of (1) may be overcome by the use of n replicated

[1] Lucio Barabesi, Dipartimento di Metodi Quantitativi, Università di Siena, P.zza S. Francesco 17, 53100 Siena, Italy.

transects randomly and independently placed on the baseline. Accordingly, let $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$, be the estimates of type (1) obtained from each replicated transect. Owing to the independent replications of the same experiment, the $\hat{X}_i$'s constitute *n iid* random variables, so that their mean $\hat{\bar{X}}$ turns out to be a well-behaved estimator for X, which is unbiased, asymptotically $(n \to \infty)$ normal with variance $\hat{\mathrm{Var}}(\hat{\bar{X}}) = \sigma^2/n$. In turn, $\mathrm{Var}(\hat{\bar{X}})$ can be consistently $(n \to \infty)$ estimated by $\hat{\mathrm{Var}}(\hat{\bar{X}}) = s^2/n$, where s^2 is the sample variance of the $\hat{X}_i$'s.

2. COMBINING INCOMPLETE SURVEYS

Now, a problem may arise if the study area is too large to be surveyed. In this case the whole area may be divided into M rectangular plots of adequate size $A_l (l = 1, 2, \ldots, M)$, in such a way that, if X_l denotes the total of marks in the l-th plot, the object parameter can be rewritten as the total of the X_l's. Accordingly, a two-stage strategy may be performed: at first, a sample G of $m < M$ plots is drawn according to an area sampling design; subsequently, in any selected plot, an estimate of the total, say $\hat{\bar{X}}_l$, is obtained on the basis of n_l replicated samples $(l \in G)$. If $p_1, p_2, \ldots, p_M$ are the first order inclusion probabilities induced by an area sampling design, a Horvitz-Thompson-like estimator for X is given by

$$\hat{\hat{X}} = \sum_{l \in G} \frac{\hat{\bar{X}}_l}{p_l}, \tag{2}$$

in which estimates of the sampled plot totals obtained from replicated transects are used instead of their true values.

It is straightforward to prove that $\hat{\hat{X}}$ constitutes an unbiased estimator for X, with variance

$$\mathrm{Var}(\hat{\hat{X}}) = \sum_{l=1}^{M} \sum_{h>l} (p_l p_h - p_{lh}) \left(\frac{X_l}{p_l} - \frac{X_h}{p_h} \right)^2 + \sum_{l=1}^{M} \frac{\sigma_l^2}{n_l p_l}, \tag{3}$$

where p_{lh} denotes the second order inclusion probability for plots l and $h (h > l)$ and σ_l^2 denotes the variance for the line transect estimator of X_l based on a single transect. Note that the first term in (3), say V_1, denotes the amount of variance due to the area sampling design, while the second term, say V_2, denotes the amount of variance due to the sampling variability of the line transect estimates in each plot.

As suggested by Brewer and Hanif (1983), if $n_l \to \infty$ for each $l \in G$, an estimator of V_1 which tends to be conservative is obtained by

$$\hat{V}_1 = \frac{M - m}{Mm} \frac{1}{m - 1} \sum_{l \in G} \left(m \frac{\hat{\bar{X}}_l}{p_l} - \hat{\hat{X}} \right)^2,$$

while an unbiased estimator of V_2 is obviously given by

$$\hat{V}_2 = \sum_{l \in G} \frac{s_l^2}{n_l p_l^2},$$

where s_l^2 denotes the sample variance of the n_l replicated estimates in the l-th plot.

As regards the design of selecting plots, the use of simple random sampling (as suggested by Jensen 1996) may prove to be a highly inefficient way. When plots vary in size, Skalsky (1994) recommends the use of a sequential sampling design in which, at each drawing, the probability of selecting a plot equals the ratio of its size to the size of the remaining plots. Note however that a certain level of homogeneity or heterogeneity is likely to be present between adjacent plots. Thus, in order to take into account the spatial structure of the data, Fattorini and Ridolfi (1997) propose a sequential sampling design in which, if a suitable factor $\beta \leq 1$ is chosen, at each drawing the probabilities of selecting those units that are adjacent to the previously selected ones are reduced or increased $(1 - \beta)$ times, according to $0 < \beta \leq 1$ or $\beta < 0$. It is at once apparent that the design exists whenever $-\infty < \beta < 1$ and it reduces to the Skalsky design when $\beta = 0$. Moreover, when $\beta = 1$, the second order inclusion probabilities vanish for contiguous units. In this case the design exists providing that, at each drawing, the remaining plots in the population are more than those which are contiguous to at least one of the previously selected plots. Finally, as $\beta \to -\infty$, the only units that can be selected are those contiguous to at least one of the previously selected units, *i.e.*, the design fails to exist if some plot has no contiguous units.

3. A COMPARISON STUDY

Some artificial populations were used to illustrate the relative precision of the two procedures. Four regions partitioned into $M = 10$ plots of unequal size were considered. As to the first region (R1), the plots in the upper part showed intensities (plot total ÷ plot size) greater than those in the lower part. As to the second region (R2), all plots had high values of intensities with the exception of a horizontal strip in the center of the study area. In the third region (R3) the intensities are approximately constant over the whole area, while as to the fourth region (R4), there were plots with high intensities that were surrounded by plots with low ones. Figure 1 contains the diagrammatic representations of the four areas together with the values of the Moran correlation coefficient ρ as an index of spatial homogeneity among the intensities of contiguous units.

The totals over the whole areas were estimated by using samples of $m = 2$ plots. Moreover it was supposed that the accuracy of the line transect estimates within each plot was proportional to the total to be estimated, *i.e.*, $\sigma_l^2 = 0.5 X_l (l = 1, 2, \ldots, 10)$, and the number of replicated samples in each plot was proportional to the size of the plot, *i.e.*, $n_l = 2A_l (l = 1, 2, \ldots, 10)$. Table 1 reports the efficiencies (variance ratio) of the Skalsky design as well as of the modified design with respect to a simple random sampling of plots. As to the modified design, since in regions R1 and R2 a spatial homogeneity occurred between contiguous units, the modified design was performed with $\beta = 1$ (*i.e.*, avoiding the selection of contiguous plots). On the other hand, since a spatial heterogeneity occurred in regions R3 and R4, the modified design was performed with $\beta \rightarrow -\infty$.

The results in Table 1 show that both the designs compare favorably with respect to simple random sampling and the modified design may provide considerable improvement with respect to the Skalsky design when an adequate level of spatial homogeneity or heterogeneity exists among contiguous units.

Table 1
Efficiencies of the Skalsky design and modified design with respect to simple random sampling for estimating the total of the areas R1, R2, R3 and R4.

Design	R1	R2	R3	R4
Skalsky	1.591	2.364	9.145	1.438
Modified	3.165	3.984	6.006	1.583

REFERENCES

Brewer, K.R.W., and Hanif, M. (1983). *Sampling with unequal probabilities*. New York: Springer-Verlag.

Buckland, S.T., Anderson, D.R., Burnham, K.P., and Laake, J.L. (1993). *Distance sampling: estimating abundance of biological populations*. London: Chapman and Hall.

Fattorini, L. (1995). Encounter sampling strategies in environmental studies. *Proceedings of the ISI Meeting on 100 Years of Sampling*, Rome, May 31-June 1.

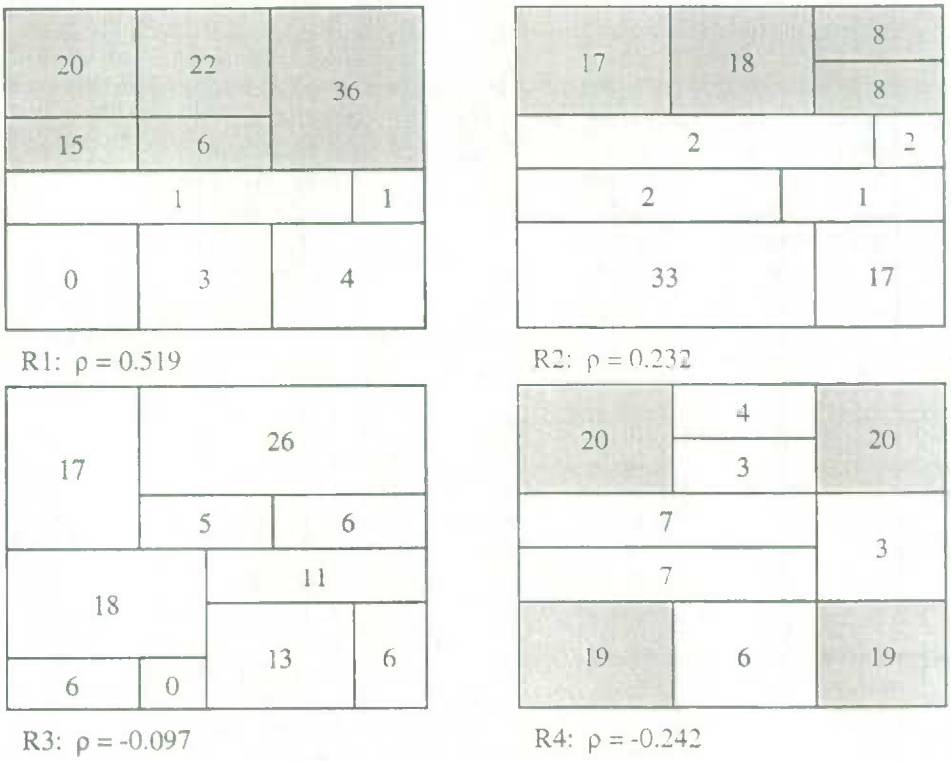
Fattorini, L., and Ridolfi, G. (1997). A sampling design for areal units based on spatial variation. *Metron LIV*, in press.

Hayne, D.W. (1949). An examination of the strip census method for estimating animal populations. *Journal of Wildlife Management*, 13, 145-157.

Jensen, A.L. (1996). Subsampling with line transects for estimation of animal abundance. *Environmetrics*, 7, 283-289.

Skalsky, J.R. (1994). Estimating wildlife populations based on incomplete area surveys. *Wildlife Society Bulletin*, 22, 192-203.

Thompson, S.K. (1992). *Sampling*. New York: John Wiley.



R1: $\rho = 0.519$

R2: $\rho = 0.232$

R3: $\rho = -0.097$

R4: $\rho = -0.242$

Figure 1. Diagrams of regions R1, R2, R3, R4.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

SAMPLE DESIGN TO INTEGRATE TWO STATEWIDE RANDOM DIGIT DIALING (RDD) SURVEYS

D. Brogan[1], D. Daniels, F. Marsteller, D. Rolka and M. Chattopadhyay

ABSTRACT

In order to estimate the prevalence of substance abuse among adolescents (12-17 years) and adults (18+ years) residing in telephone households in GA, two statistically independent random digit dialing (RDD) surveys initially were planned, one survey for each age subpopulation. This paper shows that an alternate integrated survey design, which incorporates the particular constraints of field work and number of completed interviews, required a smaller sample size of contacted households. The integrated survey design is described, along with required modified definitions for response rates. The potential of such integrated designs is discussed, with indications of the factors likely to influence their efficiency.

KEY WORDS: Random digit dialing; RDD surveys; Integrated surveys; Combined surveys; Survey efficiency; Substance abuse.

1. BACKGROUND AND SURVEY OBJECTIVES

The Center for Substance Abuse and Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA) funded all states in the U.S., as well as Washington, DC and Puerto Rico, to estimate the prevalence of substance (alcohol and drugs) abuse, substance dependence and need for treatment among adolescents and adults residing in households with residential telephone service. Point estimates were desired for each age subpopulation, further subdivided by gender, race, age and type of substance. Precision specifications for point estimates determined the required number of completed telephone interviews for adolescents and for adults.

In GA households were sampled by truncated list-assisted random digit dialing (RDD), using the Casady and Lepkowski (1993) procedure where telephone banks of size 100 with no listed residential telephone numbers within the bank are eliminated from the sampling frame. The state was stratified into four geographic areas, and within each stratum a simple random sample of telephone numbers was selected. A computer assisted telephone interview (CATI) was conducted with a randomly selected household member.

2. METHODOLOGY

2.1 Separate Surveys *vs.* An Integrated Survey

The original proposal to conduct two independent RDD surveys, one for adolescents and one for adults, had intuitive appeal for two reasons. First, each survey was a straightforward application of standard RDD sampling plans. Second, although both surveys used virtually identical data collection instruments, the additional consent procedures for adolescents made some aspects of the field work different for the two age subpopulations.

An integrated RDD survey design to cover the two age subpopulations was appealing because of lower cost. Since an estimated 15% of households contain adolescents, a survey of only adolescents will discard about 85% of households contacted and screened for adolescents. The amount of field work is nontrivial to make human contact with a household via telephone and then to screen in order to determine whether or not the household contains at least one adolescent (adolescent household). Integrating the adult and adolescent RDD surveys permits some of the nonadolescent households (households which contain no adolescents) to be used for sampling the adult subpopulation.

A practical constraint to minimize respondent burden within a household required that no more than one person per household be selected for interview. Hence, in an adolescent household, either an adolescent or an adult could be selected for interview, but not both. The number of contacted households for the integrated survey could be minimized by always selecting an adolescent for interview in an adolescent household, but this procedure would provide a biased sample of adults. Hence, it is necessary to select an adult for interview from some of the adolescent households.

2.2 Integrated Survey Design

Based on the above considerations, the design for the integrated survey is as follows. Once a household is contacted, a first stage screening question determines whether the household is adolescent or nonadolescent. If the household contains one or more adolescents, a random

[1] Donna Brogan, Professor of Biostatistics, Rollins School of Public Health, Emory University, 1518 Clifton Road N.E., Atlanta, GA 30322, U.S.A.

mechanism programmed into the CATI system immediately determines whether an adolescent or an adult is to be chosen for interview in this adolescent household, with probability p1 and (1 - *p*1), respectively. The second stage of the screening process asks for either the number of adolescents or number of adults in the household, depending upon whether an adolescent or adult is to be chosen for interview. Selection of a particular adolescent or adult for interview is by equal probability sampling, using the most recent birthday technique.

For nonadolescent households, selection of one adult for interview in all of these households generally would result in too many adults selected or termination of field work for adults earlier than for adolescents. Hence, once the first stage screening process identifies a nonadolescent household, a random mechanism within the CATI software system immediately determines whether an adult is to be selected for interview or whether the household is to be dropped from the survey and no one selected for interview, with probabilities (1 - *p*2) and *p*2, respectively. For nonadolescent households retained in the integrated survey, the second stage of the screening process determines the number of adults in the household, and the most recent birthday technique is used to select one adult for interview.

2.3 Comparing Two Separate Surveys to One Integrated Survey

The two sampling plans, separate *vs.* integrated, are compared based on the total number of households which need to be contacted. Note that the number of telephone numbers selected for either sampling plan is considerably larger than the number of contacted households, since many sample telephone numbers result in outcomes such as business, ring no answer, ring busy and nonworking number. Sample size calculations assumed the following:

1. 15% of households contain at least one adolescent, based on U.S. Census data for GA.
2. Every household contains at least one adult.
3. 90% of contacted households participate in the entire screening process.
4. 78% of persons (adolescents or adults) selected for interview are interviewed.

2.4 Plan 1 (6900 Adult Interviews and 3400 Adolescent Interviews)

The initial sample size, based on precision and cost, was completed interviews for 6,900 adults and 3,400 adolescents. An RDD survey of adults would require 6,900 / [.90 × .78] = 9,829 contacted households, and an RDD survey of adolescents would require 3,400 / [.15 × .90 × .78] = 32,289 contacted households. Thus, the total number of households which need to be contacted for two independent surveys is 32,289 + 9,829 = 42,118.

In planning the integrated survey, initial investigation of possible choices for p1 and p2 did not reveal an easily obtainable optimum solution which would minimize the total number of contacted households; this analytical work is in process. Our strategy was to make a reasonable choice for p1, which then leads to the value for p2 being determined by the fixed number of completed interviews among adolescents and adults.

Selection of p1 = 1.0, *i.e.*, always choose an adolescent for interview in an adolescent household, minimizes the required number of contacted households for the integrated survey but would produce a biased sample of adults. Selection of p1 close to 1.0 increases the variability of sampling weights among adults, as shown by the following ratio of the first two components of the initial weight for selected adults in adolescent households compared to selected adults in nonadolescent households: [(.85) (1 -p2)] / [(.15) (1-p1)]. This increased variability in the sampling weights yields increased variance of point estimates for adults, but the substantial variation already in the sampling weights for adults due to other factors may reduce somewhat the impact of choice of p1 on variability of sampling weights in the adult subpopulation.

After investigating several choices for p1, we chose p1 = .87 as the probability an adolescent is chosen for interview in an adolescent household, with an adult chosen if an adolescent is not. For p1 = .87, p2 was determined to be .71136. Thus, among contacted nonadolescent households, a household was dropped from the survey with probability .71136 and retained with probability .28864, with one adult being selected for interview in each retained household. The required sample size for the integrated survey under plan 1 is driven by the required number of adolescent interviews. Hence, the required number of contacted households is calculated as 3,400 / [.15 × .90 × .87 × .78] = 37,113. Just to check the calculation for p2 as .71136, the number of adult interviews expected with 37,113 contacted households is given by:

$$37{,}113 \times [(.15 \times .90 \times .13 \times .78) +$$

$$(.85 \times .90 \times .28864 \times .78)] = 508 + 6{,}392 = 6{,}900,$$

with 508 and 6,392 completed adult interviews from adolescent and nonadolescent households, respectively.

The approach of two independent surveys requires a sample size of 42,118 contacted households, whereas the integrated survey requires a sample size of 37,113 contacted households. The ratio of the two sample sizes is 42,118/ 37,113 = 1,135, indicating that the design of two independent surveys requires a 13.5% larger sample size than the integrated design.

Note that a comparison of two independent surveys to the integrated survey based only on number of contacted households makes several assumptions, which may or may not be true in practice. First, the cost and response rates are assumed to be constant over several different household screening combinations, some of which are more time consuming than others. Second, the interview response rate is assumed to be the same for adults as for adolescents. This seemed reasonable in our survey but may not be true in general. Third, if the actual percentage of households which are adolescent differs significantly from the assumed 15%, this will impact the comparison of the two sampling plans.

2.5 Plan 2 (6000 Adult Interviews and 2700 Adolescent Interviews)

Based on the calculations for Plan 1, we decided to conduct an integrated RDD survey. However, subsequent negotiations indicated that a sample size of 6,900 adult interviews and 3,400 adolescent interviews was too large for the fixed budget. Hence, the sample size was reduced to completed interviews for 6,000 adults and for 2,700 adolescents. Conducting two independent RDD surveys for the completed interview sizes specified in plan 2 requires a total of 34,188 contacted households, *i.e.*, 6000/[(.90) × (.78)] = 8547 contacted households for the adult survey and 2700/[(.90) × (.15) × (.78)] = 25,641 contacted households for the adolescent survey.

For the integrated RDD survey in plan 2, we chose p1 = .79 and (1 - p1) = .21 for the probabilities of selecting an adolescent or adult, respectively, from an adolescent household. We increased the probability (1 - p1) to .21 from the .13 used in plan 1 because it seemed that the undersampling of adults in adolescent households might be too extreme with .13. The choice of p1 = .79 determined that p2 = .727255, *i.e.*, a probability of .273 to retain a nonadolescent household in the sample and select one adult for interview. With the integrated survey design for plan 2, the probability that a contacted household results in a completed adolescent interview is (.90) × (.15) × (.79) × (.78) = .083187. Similarly, the probability that a contacted household results in a completed adult interview is [(.90) x (.15) × (.21) × (.78) + (.90) × (.85) × (.272745) × (.78)] = .1848599. Hence, the required sample size of contacted households is 32,457, given by either the adolescent or adult sample size requirement, *i.e.*, 32,457 = 2,700/(.083187) = 6,000/(.1848599).

For plan 2, the ratio of the required number of contacted households for two independent surveys versus the integrated survey is 34,188/32,457 = 1.053. Hence, conducting two independent surveys requires a 5.3% increase in the number of contacted households, compared to the integrated survey. This is a smaller increase than the 13.5% in plan 1, indicating that the integrated survey is not as efficient in plan 2 as in plan 1. However, the integrated design is still better than two independent surveys under both sample size plans. The decreased value of p1 in plan 2 makes it harder for the integrated survey to be dramatically better than two independent surveys.

2.6 Response Rate Definitions for the Integrated Survey

Typical response rate definitions used in RDD surveys require modification for an integrated survey design. For example, some nonadolescent households are not retained for sampling adults, and these households should not count against the response rate. Also, an overall response rate is desired for each subpopulation (adults and adolescents). Four stage-specific response rates are defined below for the integrated survey. The first two rates, contact and screen1, are the same for the adolescent and adult subpopulations. The next two rates, Screen 2 and completion, are specific to the adult or adolescent component of the integrated survey. The overall survey response rate for each subpopulation is the product of the relevant four stage-specific rates. Due to space limitations, detailed formulas are not given for the stage-specific response rates.

- Contact Rate, defined as proportion of residential telephone numbers where human contact is made with a household member (*i.e.*, someone answers the telephone).
- Screen 1: Rate, defined as proportion of contacted residential telephone numbers where screening information is obtained to classify the household as adolescent or nonadolescent.
- Screen 2: Rate, defined as proportion of households designated for selection of an adolescent (adult) for interview where screening information is obtained on the number of adolescents (adults) so that a particular adolescent (adult) can be selected for interview.
- Completion Rate, defined as proportion of eligible adolescents (adults) selected for interview who complete an interview.

3. RESULTS OF FIELD WORK

In the field work 70,923 telephone numbers were used, of which 41,415 were working residential numbers and 37,679 were contacted households, for a contact rate of 37,679/41,415 = .9098. The Screen1 Rate was 31,738/37,679 = .8423, and the Screen 2 Rate was 4,591/4,861 = .9445 for adolescents and 9,146/9,624 = .9503 for adults. The completion rate was 3,493/4,072 = .8578 for adolescents and 7,713/8,519 = .9054 for adults. Hence, the overall survey response rate was .6209 for adolescents and .6593 for adults. We had anticipated an overall survey response rate of .70, *i.e.*, (.90) × (.78). The lower than anticipated response rates were due primarily to the first two stages, *i.e.*, contact and screen 1; the product of these two rates alone was .7663.

The proportion of households with at least one adolescent was estimated from field work data as (6,391)/(6,391 + 25,347) = .20, as compared with our planning figure of .15. The proportion of adolescent households which were selected for an adolescent interview was 4,861/6,391 = .76, a slight deviation from the planned figure of p1 = .79. The proportion of nonadolescent households in which an adult was selected for interview was 8,094/25,347 = .319, larger than the survey design figure of (1 - p2) = .273. In fact, the values of p1 and p2 were changed three times during field work to correct some programming errors and to reflect the larger (than anticipated) proportion of households which contained at least one adolescent. A major advantage of the integrated survey design implemented via CATI is the ability to change the design probabilities p1 and p2 during field work, as needed. Of course, the actual design probabilities which are used for each household need to be recorded so that appropriate sample weights can be calculated later.

In summary there were 37,679 households contacted with 3,493 interviewed adolescents and 7,713 interviewed adults. Plan 2 specified contact with about 32,457 households, with 2700 and 6000 adolescent and adult interviews, respectively. There were 29.4% more adolescent interviews and 28.6% more adult interviews than planned, and contact with 16.1% more households than planned. Two reasons contributed to the excess number of completed interviews and contacted households: (1) initial incorrect specification of number of completed interviews required for the two strata in south GA and (2) field work protocol to complete all work within a block of telephone numbers released into the sample.

4. DISCUSSION

Based on the one survey discussed here, it probably was more expensive to conduct the integrated survey rather than two independent surveys. However, the reasons are related to survey implementation rather than survey design. In general, the concept of an integrated RDD survey design in situations like the one described here has clear potential. In the two specific plans presented, the strategy of two independent RDD surveys, compared to an RDD integrated survey, increased the required sample size of contacted households by 5.3% to 13.5%. The characteristics and efficiency of the integrated survey design depend heavily on the choice made for the probability p1 and the resulting determination of p2. Based on the limited experience in this one survey, however, it seems that an integrated survey may not always be practically more efficient than two independent surveys. Current work is underway to investigate the impact of the following parameters on the efficiency of the integrated survey, compared to two independent surveys:

- the ratio of completed interviews required for the more common subpopulation and for the rarer subpopulation (in this case adults to adolescents, around 2.0);
- the ratio of the prevalence (on a household basis) of the more common subpopulation to the rarer population (in this case about 1.0/0.15 or 1.0/0.20, *i.e.*, about 5 to 7);
- any differential costs of household screening procedures in the integrated survey versus two independent surveys;
- any differential response rates in the two subpopulations for screening or interviewing;

An important realistic constraint in our survey was the decision to select no more than one person per household for interview because of respondent burden. An integrated design that allows an adolescent interview in every adolescent household, with the possibility of also interviewing an adult in that household, will be more efficient than both of the integrated plans we considered. Further, interviewing both an adult and an adolescent within the same household might allow interesting analyses of the correlation of substance abuse and dependence among adolescents and adults within the same household. For example, with the number of completed interviews for plan 1, an integrated survey could be done with 32,289 contacted households, where each household has probability .3044 of being designated for an adult interview and every adolescent household is designated for an adolescent interview. Under this scenario two independent surveys would require 30.4% more contacted households than the integrated survey. The proportion of households where two persons are selected for interview is .0457, and the proportion of adolescent households where an adult is also selected for interview is .3044.

RDD sampling of the general U.S. population or specific subpopulations has been preferred for many years over area probability sampling because of its assumed lower cost, primarily resulting from the avoidance of significant travel and time costs to count, list and sample housing units in the field. Further, if personal interviewing is done on the telephone, rather than at the subject's home or office, further savings are realized on interviewers' travel and time costs. Telephone surveys today typically use the telephone both for sampling and for interviewing, expecting this approach to be much cheaper than other alternatives.

However, several recent phenomena have coalesced to increase the costs of telephone surveys, perhaps dramatically. These include proliferation of telephone numbers which are not residences (*e.g.*, fax machines, cellular phones, computer lines), area codes which no longer designate a unique geographic area, people not at home to answer the telephone and refusal to answer the telephone even if they are at home, and declining response rates to telephone surveys. Significantly more resources and money are needed now to achieve RDD response rates equivalent to that of the past. In this era of increasing costs of conducting RDD surveys, the option of integrating RDD surveys has even more appeal.

ACKNOWLEDGMENT

This research was conducted in conjunction with the Georgia Department of Human Resources, the Psychiatry Department of the School of Medicine at Emory University, the Biostatistics Department of the Rollins School of Public Health at Emory University and the Gallup Corporation. This study was made possible through funding from CSAT contract number 270-94-0010, State Demand and Needs Assessment Studies: Alcohol and Other Drugs.

REFERENCES

Casady, R.J., and Lepkowski, J.M. (1993). Stratified telephone survey designs. *Survey Methodology*, 19, 103-113.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

SYMBOLIC COMPUTATION OF BOOTSTRAP VARIANCE ESTIMATORS IN SAMPLE SURVEY THEORY

G.T. Sampson[1]

ABSTRACT

The bootstrap was introduced in 1979 as a computer-intensive method for estimating the standard error and for studying other properties of an estimator. The same objective can be achieved, in many cases, by the development and implementation of computer algorithms for the symbolic computation and evaluation of Taylor series expansions. The analytic bootstrap permits the calculation of standard errors for a wide variety of estimators in a fraction of the time and with comparable or better accuracy when compared with conventional bootstrap Monte Carlo resampling. In addition, properties such as bias and variance of bootstrap estimators may be assessed. The methodology is illustrated here through an application to the analytic bootstrap variance calculation for the estimator of a ratio under the special case of simple random sampling with replacement. These calculations can now be performed instantaneously on a computer without human error and without the simulation error associated with conventional bootstrap techniques. Moreover, calculations can be performed where no formulae presently exist. Under more complex survey designs, the method has the potential to afford statistical agencies savings with respect to surveys that presently utilize resampling methods for variance estimation.

KEY WORDS: Computer algebra; Linearization; Ratio estimator; Resampling methods; Survey estimators; Variance estimation.

1. INTRODUCTION

In 1979 Efron proposed the idea of using computer-based simulations instead of mathematical calculations to obtain the sampling properties of estimators $\hat{\theta}$. This paper focuses on how the same objective can be achieved, in many cases, by the development and implementation of computer algorithms for the symbolic computation and evaluation of asymptotic expansions. Our discussion is largely confined to the situation in which $\hat{\theta}$ can be expressed as a smooth function of means or totals of independent and identically distributed (*iid*) random variables. The corresponding class of statistics is wide and includes, for example, sample means, variances, ratios and differences of these, correlations, maximum likelihood estimators and *M*-estimators. But it is by no means comprehensive. It does not include, for example, the sample median or other order statistics; see Sampson (1997) for an alternative treatment for non-smooth statistics.

Much of sampling theory as well as other areas of statistics can involve cumbersome algebra. In many cases, the underlying structure of this algebra exhibits a repetitive pattern. If a particular pattern can be identified in the form of a fundamental rule, the possibility of automating a calculation arises. Seemingly unrelated formulae can result from the same fundamental rule and one computer algebra tool can be constructive in implementing many calculations. Using this approach, Stafford and Bellhouse (1997) presented the basic building blocks to develop a computer algebra for survey sampling theory. They showed that three basic techniques in sampling theory depend on the repeated application of several fundamental rules. Their methodology permits the symbolic computation of moments of typical survey estimators as well as the determination of unbiased estimators under simple random sampling without replacement. This has recently been extended to include other sampling designs. Similarly, Sampson (1997) developed symbolic computational tools for the derivation of nonparametric bootstrap moments for statistics belonging to the class of *M*-estimators.

This article investigates the automatic derivation of analytic bootstrap variance estimators associated with typical survey estimators under the special case of *iid* random variables, or simple random sampling with replacement. These analytic calculations can now be performed instantaneously on a computer without human error and without the simulation error inherent in Monte Carlo techniques. A great deal of hand-written algebra can be avoided. Moreover, calculations can be performed where no formulae presently exist. Under more complex survey designs, the method has the potential to afford statistical agencies savings with respect to surveys, such as Statistics Canada's National Population Health Survey and the Labour Force Survey, that presently utilize resampling methods for variance estimation. For this reason, it may be useful to incorporate the technology discussed here into an integrated computer package for estimation in survey sampling, see Hidiroglou *et al.* (1997). The computer code that was used to generate the analytic expressions in this paper is based on Andrews, Stafford and Wang (1993). It was written in the symbolic package *Mathematica 3.0* created by Wolfram (1996) and is available from the author.

[1] George T. Sampson, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; e-mail: sampgeo@statcan.ca.

Section 2 reviews conventional bootstrap estimation of variance. Automation of analytic bootstrap calculations is discussed in Section 3. Section 4 presents several *Mathematica* operators that can be used to automatically generate analytic bootstrap variance estimators and other expressions in the *iid* case, and then the methodology is applied to the case of estimating a ratio. Finally, a summary and a discussion of future work completes Section 5.

2. REVIEW OF MONTE CARLO BOOTSTRAP ESTIMATION OF VARIANCE

Suppose that the data $X = (x_1, ..., x_n)$ is an observed random sample of size n from a population with distribution function F. Let $\hat\theta = \hat\theta(X)$ be a statistic of interest. Then the variance of $\hat\theta$ is defined as

$$\mathrm{Var}_F(\hat\theta) = E_F[\hat\theta - E_F(\hat\theta)]^2. \qquad (2.1)$$

The bootstrap, originally developed by Efron (1979), is a combination of two techniques. We review these as follows.

(1) **The substitution principle**: We estimate (2.1) by replacing F by its nonparametric estimator $F = \hat F$ corresponding to the empirical distribution function, F_n, defined to be the discrete distribution that assigns probability $1/n$ to each value of x_i for $i = 1, ..., n$. This provides the theoretical form of the bootstrap variance estimator of $\hat\theta$, say, $v_{\mathrm{boot}} = \mathrm{Var}_{\hat F}(\hat\theta)$. It may be used directly for practical applications only when v_{boot} is an explicit function of $(x_1, ..., x_n)$ in which case its evaluation *is equivalent to replacing expectations by sample averages in* (2.1). For example, when $\hat\theta = \bar x$ then (2.1) becomes

$$\mathrm{Var}_F(\bar x) = \mathrm{Var}_F(x) \cdot \frac{1}{n} = E_F[x - E_F(x)]^2 \cdot \frac{1}{n} \qquad (2.2)$$

for which the substitution $E_{\hat F}(\cdot) \rightarrow \frac{\sum(\cdot)}{n}$ in (2.2) gives

$$v_{\mathrm{boot}} = \mathrm{Var}_{\hat F}(\bar x) = \frac{\sum_{i=1}^{n} (x_i - \bar x)^2}{n} \cdot \frac{1}{n} = \frac{(n-1)}{n} \cdot \frac{s^2}{n}. \qquad (2.3)$$

However, in most cases $\hat\theta$ is complicated and v_{boot} is then approximated by Monte Carlo as follows.

(2) **Numerical approximation**: We draw B bootstrap samples $(x^*_{1b}, ..., x^*_{nb})$, $b = 1, ..., B$, independently and with replacement from $(x_1, ..., x_n)$, compute $\hat\theta^*_b$ and approximate v_{boot} by computing the empirical variance of the bootstrap replicates $\hat\theta^*_b$.

Andrews and Stafford (1993) developed tools using the computer algebra package *Mathematica* for the symbolic computation of asymptotic expansions of many common statistics and their moments. This was further extended to include a wide class of statistics known as *M*-estimators (solutions to estimating equations) in Andrews and Feuerverger (1993) as described in Sampson (1997) and Yun (1998). Most statistics in this class, including typical survey estimators, are smooth functions of sample means or totals. Computer algebra algorithms permit expression of such estimators $\hat\theta$, either exactly or approximately, in terms of sums of products of means or totals. Moreover, moments and cumulants of $\hat\theta$, such as (2.1), may also be represented as power series expansions which are valid for *any* distribution F. Finally, nonparametric bootstrap moments or bootstrap cumulants of $\hat\theta$, such as v_{boot}, assume the empirical distribution $F = \hat F$ and are directly available by replacing simple expectations by sample averages as described in step (1) and the resampling process in step (2) is entirely avoided.

3. AUTOMATION OF ANALYTIC BOOTSTRAP CALCULATIONS

3.1 Introduction

The derivation of analytic bootstrap variances or bootstrap expectations of smooth functions of sample means or totals involves the following three steps:

(1) Express $\hat\theta$, either exactly if $\hat\theta$ is linear or approximately if $\hat\theta$ is nonlinear, as the sum of products of sample means or totals.

(2) Express the expectation or variance of $\hat\theta$ as a power series to a desired order.

(3) Obtain the bootstrap expectation or bootstrap variance of $\hat\theta$ by replacing expectations by sample averages in (2).

These steps can provide analytical approximations to the theoretical bootstrap variance (or expectation) in the case of nonlinear estimators. For a statistician working by hand, this would be a simple, mechanical but laborious task. However, as described above, recent advances in applications of computer algebra to statistics have succeeded in automating much of the algebra underlying the mathematical and statistical procedures referred to here. The immediate benefit to the statistician is that calculations can now be performed instantly on the computer without error and to any order of accuracy desired. We describe below one of several techniques that has been used in the process of automating steps (1) to (3) above. A full exposition of these techniques is given in Stafford and Bellhouse (1997) in the context of simple random sampling without replacement.

3.2 Analytic Bootstrap Expectation of an Estimator

A statistic $\hat\theta$, expressible as a smooth function of means, may be written as an asymptotic expansion where terms descend in order by $1/\sqrt{n}$, specifically

$$\hat\theta = \hat\theta_0 + \hat\theta_1/\sqrt{n} + \hat\theta_2/n + ... \qquad (3.1)$$

where $\hat\theta_i$ is the coefficient of the $n^{-i/2}$ term and is a linear combination of products of exactly i centered, normalized sums Z_{ij}, $j = 1, ..., i$, of *iid* random variables, see Section 4 for illustration. The quantities Z_{ij} are $O_p(1)$ random variables. Hence, (3.1) essentially represents a sum of products of the sums Z_{ij}. The whole operation of finding the expectation of (3.1) may be represented schematically as $\sum\prod\to\sum\sum\to\sum\prod$ where $\sum\prod$ denotes a sum of products and $\sum\sum$ denotes a sum of nested (*i.e.*, disjoint) sums. We illustrate this schema for the simple case of finding $E_{\text{boot}}(\bar{x}^2)$ under simple random sampling with replacement. We first observe that the product of two sums may be decomposed into a sum of (nested) sums over disjoint indices as in

$$\sum_{i=1}^{n} x_i \cdot \sum_{j=1}^{n} x_j = \sum_{i=1}^{n} x_i^2 + \sum_{i\neq j} x_i x_j \tag{3.2}$$

This is the $\sum\prod\to\sum\sum$ step. Now the expectation operator is applied to $\sum\sum$. Since $E(x_i \cdot x_j) = E(x_i)\cdot E(x_j)$ for $i \neq j$, due to independence, then the expectation of both sides of (3.2) yields

$$\begin{aligned} &E\left(\sum_{i=1}^{n} x_i \cdot \sum_{j=1}^{n} x_j\right) \\ &= \sum_{i=1}^{n} E\left(x_i^2\right) + \sum_{i\neq j} E(x_i)\cdot E(x_j) \\ &= \sum_{i=1}^{n} E\left(x_i^2\right) + \left[\sum_i E(x_i)\cdot \sum_j E(x_j) - \sum_i \left[E(x_i)\right]^2\right] \\ &= n\cdot E(x^2) + n^2\cdot [E(x)]^2 - n\cdot [E(x)]^2 \end{aligned} \tag{3.3}$$

which gives

$$E\left(\bar{x}^2\right) = \frac{E(x^2)}{n} + [E(x)]^2 - \frac{[E(x)]^2}{n}. \tag{3.4}$$

Note that the middle step in (3.3) involved the substitution $\sum\sum\to\sum\prod$, that is, the nested sum in (3.2) was re-expressed as a linear combination of simple products of sums *after* having applied the expectation operator. Finally, replacing expectations by averages in (3.4), we obtain the formula for the bootstrap second moment of $\bar{x}$, given by

$$E_{\text{boot}}\left(\bar{x}^2\right) = \frac{\sum_{i=1}^{n} x_i^2}{n^2} + \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^3} \tag{3.5}$$

In general, $\text{Var}_{\text{boot}}(\hat{\theta})$ may be derived analytically using similar steps. Indeed, any smooth function $g(\hat{\theta})$, such as a power of $\hat{\theta}$ or $\hat{\theta} - E(\hat{\theta})$, reduces algebraically to a series expansion similar in structure to (3.1) to which the expectation operator may then be applied.

The identification of terms in a Taylor series expansion of an estimator as well as the elementary operations associated with the process $\sum\prod\to\sum\sum\to\sum\prod$ for the evaluation of an expected value can be automated by generating the corresponding underlying algebraic structures recursively using several fundamental rules programmed in *Mathematica*. For a detailed account see Stafford and Bellhouse (1997). The analytic expressions illustrated in the next section have been symbolically derived using similar tools.

4. ESTIMATOR OF A RATIO

In this section we present several symbolic computational operators that may be used in the context of providing analytic bootstrap variance expressions for common or uncommon survey estimators, automatically. We illustrate these techniques for the case of the estimator of a ratio.

The estimator of a ratio of population means, $R = E(y)/E(x)$, is given by $\hat{R} = \bar{y}/\bar{x}$, where $\bar{y}$ and $\bar{x}$ represent sample means corresponding to the measurement of interest and the auxiliary information, respectively. Then $\hat{R}$ is a product of two quantities, $\bar{y}$ and $1/\bar{x}$, each having an asymptotic expansion of its own. In general, for the purpose of making asymptotic expansions, we define an $O_p(1)$ random variable for an estimator $\hat{\theta}$ given by $Z(\hat{\theta}) = \sqrt{n}\left[\hat{\theta} - E(\hat{\theta})\right]$. Hence, the expansion for $\bar{y}$ is $E(y) + Z(\bar{y})/\sqrt{n}$, exactly. The expansion for $1/\bar{x}$ results from a similar construction and then applying a Taylor expansion to $[E(x) + Z(\bar{x})/\sqrt{n}]^{-1}$. Note that this quantity may be re-expressed as

$$\left[E(x) + \frac{Z(\bar{x})}{\sqrt{n}}\right]^{-1} = \frac{1}{E(x)\left[1 + \dfrac{Z(\bar{x})}{E(x)\cdot\sqrt{n}}\right]} \tag{4.1}$$

which can then be linearized by using $(1+\varepsilon)^{-1} = 1 - \varepsilon + \varepsilon^2 - \ldots$. Hence, $\hat{R}$ can be expressed as a series expansion in terms of the centered, normalized sample means $Z(\bar{x})$ and $Z(\bar{y})$.

In general, we shall define a given survey estimator in terms of $O_p(1)$ random variables, as illustrated above, prior to applying the symbolic functions described next. We introduce five basic computer algebra operators as follows:

1. `Taylor`[estimator, i] expresses an estimator as a sum of products of sample means correct to order *i*, using Taylor linearization when necessary.
2. `Expt`[estimator, i] returns the expected value of $\hat{\theta}$ as a series expansion correct to order *i*.
3. `Var`[estimator, i] returns the variance of $\hat{\theta}$ as a series expansion correct to order *i*.
4. `BootExp`[estimator, i] returns the expansion in step 2 with expected values replaced by averages.
5. `BootVar`[estimator, i] returns the expansion in step 3 with expected values replaced by averages.

For example, the series expansion of the variance of $\hat{R}$ correct to order n^{-2} is given by

`Var`[ratio,4] =

$$\left(\frac{E[X^2]E[Y]^2}{E[X]^4} - \frac{2E[Y]E[XY]}{E[X]^3} + \frac{E[Y^2]}{E[X]^2}\right)/n +$$
$$(-\frac{E[X^2]E[Y]^2}{E[X]^4}) + \frac{8E[X^2]^2E(Y)^2}{E[X]^6} - \frac{2E[X^3]E[Y]^2}{E[X]^5} +$$
$$\frac{2E[Y]E[XY]}{E[X]^3} - \frac{16E[X^2]\,E[Y]E(XY]}{E[X]^5} + \frac{5E[XY]^2}{E[X]^4} +$$
$$\frac{4E[Y]E[X^2Y]}{E[X]^4} - \frac{E[Y^2]}{E[X]^2} + \frac{3E[X^2]E[Y^2]}{E[X]^4} -$$
$$\frac{2E[XY^2]}{E[X]^3})/n^2. \qquad (4.2)$$

Note that the leading $1/n$ term in (4.2) reproduces the usual Taylor linearization variance (see, for example, Särndal *et al.* (1992, p.179), Binder (1983)). Finally, the fourth order series expansion of the bootstrap variance of $\hat{R}$ is obtained as

BootVar[ratio,4] =

$$\left(\frac{n\,\text{sum}[X^2]\,\text{sum}[Y]^2}{\text{sum}[X]^4} - \frac{2n\,\text{sum}[Y]\,\text{sum}[XY]}{\text{sum}[X]^3} + \frac{n\,\text{sum}[Y^2]}{\text{sum}[X]^2}\right)/n -$$
$$(-\frac{n\,\text{sum}[X^2]\,\text{sum}[Y]^2}{\text{sum}[X]^4}) + \frac{8n^2\text{sum}[X^2]^2\text{sum}(Y)^2}{\text{sum}[X]^6} -$$
$$\frac{2n^2\text{sum}[X^3]\,\text{sum}[Y]^2}{\text{sum}[X]^5} + \frac{2n\,\text{sum}[Y]\,\text{sum}[XY]}{\text{sum}[X]^3} -$$
$$\frac{16n^2\text{sum}[X^2]\,\text{sum}[Y]\,\text{sum}[XY]}{\text{sum}[X]^5} + \frac{5n^2\text{sum}[XY]^2}{\text{sum}[X]^4} +$$
$$\frac{4n^2\text{sum}[Y]\,\text{sum}[X^2Y]}{\text{sum}[X]^4} - \frac{n\,\text{sum}[Y^2]}{\text{sum}[X]^2} +$$
$$\frac{3n^2\text{sum}[X^2]\,\text{sum}[Y^2]}{\text{sum}[X]^4} - \frac{2n^2\text{sum}[XY^2]}{\text{sum}[X]^3})/n^2 \quad (4.3)$$

where, for example,

$$\text{sum}\ [XY]^2 = \left[\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j\right]^2.$$

Observe that (4.2) and (4.3) are analytical formulas. Moreover, expression (4.3) is an estimator which may be easily evaluated for any given sample of *iid* bivariate data. It represents an analytical approximation to the theoretical bootstrap variance of $\hat{R}$ to terms of order n^{-2}.

We have presented 4th order expressions here although series expansions to *any* order of accuracy may be generated automatically by using the operators described above. Note that `BootVar` $[\bar{x}, 2]$ gives (2.3) and `BootExp` $[\bar{x}^2, 4]$ reproduces (3.5). For a comparison between symbolic bootstrap variance estimates involving the correlation coefficient and results obtained from conventional bootstrap Monte Carlo resampling see Sampson (1997).

5. SUMMARY AND DISCUSSION

This paper has presented a methodology that may be used to derive analytic bootstrap variance estimators and related expressions for typical survey estimators automatically under the special case of simple random sampling with replacement from an infinite population. Monte Carlo resampling is entirely avoided. The resulting bootstrap expressions are formulas which may be efficiently evaluated for given data. The proposed technology permits labour-free derivation of analytical approximations to theoretical bootstrap variances and expectations to any order of accuracy desired. The method has the potential to be cheaper than conventional bootstrap techniques.

The next phase in this work is to extend the *iid* unistage results to stratified simple random sampling with replacement. Rao and Wu (1988) demonstrated that naïve bootstrap variance estimators are not consistent estimators of variances of general nonlinear statistics in the context of stratified random sampling when the strata sample sizes n_h are small. They solved this scaling problem by showing that, for example, the use of bootstrap strata sample sizes of $n_h - 1$ instead of n_h leads to consistent bootstrap estimators. Future work will consider using this approach to provide analytical approximations to the theoretical bootstrap variance using symbolic computation.

ACKNOWLEDGEMENTS

The author is grateful to J.N.K. Rao and to Wesley Yung for their constructive suggestions and helpful comments in the preparation of this paper.

REFERENCES

Andrews, D.F., and Feuerverger, A. (1993). General saddlepoint approximation methods for bootstrap configurations. University of Toronto Technical Report.

Andrews, D.F., and Stafford, J.E. (1993). Tools for the symbolic computation of asymptotic expansions. *Journal of the Royal Statistical Society*. (B), 55, 613-628.

Andrews, D.F., Stafford, J.E., and Wang, Y. (1993). ASTAT: a system for calculating asymptotic expansions. University of Toronto Technical Report.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Hidiroglou, M.A., Bellhouse, D.A., and Stafford, J.E. (1997). Generalized estimation systems with future enhancements using symbolic computation. Statistics Canada Technical Report.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Sampson, G.T. (1997). Symbolic computation of nonparametric bootstrap estimators and their properties. Ph.D. Thesis, University of Toronto.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Stafford, J.E., and Bellhouse, D.R. (1997). A computer algebra for sample survey theory. *Survey Methodology*, 23, 3-10.

Wolfram, S. (1996). *The Mathematica Book* (3rd Edition). Wolfram Media and Cambridge University Press.

Yun, S. (1998). The efficient hybrid bootstrap estimator: Exploiting unbiasedness of Monte Carlo and efficiency of symbolic computation. Ph.D. Thesis, University of Toronto.

SESSION C-4

Dealing with Non-Response

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

RESPONSE ANALYSIS SURVEYS: WHAT ARE THEY AND WHAT CAN WE GAIN FROM THEM? A CASE STUDY USING THE SURVEY OF FAMILY EXPENDITURES

W. Rea and M. Singh[1]

ABSTRACT

Response analysis surveys provide a relatively inexpensive technique for developing or evaluating the data collection portion of the survey process. For interviewer-conducted surveys, they can be used to explore the cognitive processes of the respondent before, during, and after the interview. In particular, they can provide information about the respondents' reactions to introductory letters/brochures, their relationships with the interviewer, their attitudes to the survey, their reasons for participating, *etc.* They also allow respondents to give feedback on the survey process and suggest improvements. The Household Surveys Division of Statistics Canada tested the usefulness of a response analysis survey on respondents to the interviewer-conducted Family Expenditures Survey (FAMEX). The purpose of the test was to explore respondent reaction to the FAMEX data collection process with a view to increasing the survey response rate and also to determine whether a response analysis survey is a useful source of information about the attitudes of respondents. Respondents' motivation for participating, their perception of the purpose of the survey and the uses of the data, and their reaction to the way this information was presented to them are analyzed and presented in the paper. It is concluded that a well-planned response analysis survey should lead to improved respondent relations and, hopefully, an improved response rate.

KEY WORDS: Response analysis survey; Feedback questionnaire; Family expenditure surveys; Introductory materials.

1. THE FAMEX SITUATION

The Family Expenditure (FAMEX) Survey at Statistics Canada (STC) collects information on household expenditure and income for a one year period. The survey has been conducted about once every four years. FAMEX is often referred to as a "recall" survey, because respondents are asked to report their expenditures over a period of time, using memory supplemented by records. The expenditure categories are detailed, resulting in an interview that lasts 2.5 to 3 hours on average.

Concern has been expressed about the extent of the "burden" placed on the respondent in terms of time required, invasion of privacy, and the difficulty of recalling so many expenditures. The decline in response rate for the survey from 81.4% in 1982 to 73.8% in 1992 has only increased this concern. To assist in improving response rates the survey was deemed mandatory (participation was required, by law) for the 1996 survey cycle.

In addition, the survey will soon become an annual survey and the sample size will be increased to satisfy new reliability requirements for provincial estimates. Due to these new requirements, increase in survey frequency and sample size, there is a need to become more proactive with respect to response burden.

With this in mind, a project based on the 1996 FAMEX survey was initiated to examine respondent relations issues such as respondent reaction to the FAMEX survey and its effect on response rates. The implementation of a response analysis survey (referred to in this paper as the feedback survey) was one component of this project and will be the focus of this paper. The other part consisted of focus groups with both respondents and interviewers that were conducted across the country to collect qualitative information about respondent reaction to the survey experience and suggestions for improving it.

1.1 Response Analysis Surveys

Response analysis surveys (RASs) are a relatively low-cost tool for developing or evaluating the data collection portion of the survey process. Typically, they are used in the questionnaire development stage of survey design to improve the wording and layout of questionnaires. They may complement other questionnaire evaluation procedures such as one-on-one interviewing and focus group testing. Essentially, response analysis surveys provide a respondent debriefing about the survey – a retrospective analysis conducted after a respondent has completed the main survey (Goldenberg *et al.* 1993). Most of the literature cites cases where the response analysis survey is a telephone administered survey about a mail out/mail back survey, usually of a business. Response analysis surveys are, thus, "surveys about surveys" (Bercini 1991) conducted using a methodology different from that of the survey being evaluated.

The feedback survey used to evaluate the FAMEX survey is different from those in previous studies in the following ways:

- the collection unit is a household and not a business;
- the survey being evaluated is interviewer assisted and not mail out/mail back;

[1] Mamta Singh, Statictics Canada, 5th Floor, Section A1, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

- the response analysis survey is a pre-paid postage, paper survey and not telephone administered; and
- the FAMEX feedback survey collected information that could improve not only the FAMEX questionnaire but also aspects of the FAMEX data collection process.

2. SCOPE AND LIMITATIONS

The scope of this project was to conduct a response analysis survey to:

- learn more about respondent reaction to the FAMEX survey in order to reduce response burden and improve response rates; and
- determine whether feedback questionnaires provide useful (cost effective, relevant) qualitative information about respondent attitudes.

In defining the scope of this project certain constraints were identified:

- This project was designed to be a test of the response analysis survey technique and as a test was only conducted in two out of the six regional offices in Canada.
- No follow-up of respondents was undertaken to improve the feedback survey response rate.
- We assumed that we received a large number of negative comments because of the nature of a feedback questionnaire (asks for criticism), due to the mandatory nature of the FAMEX survey, and because many of the feedback questionnaires were completed by converted "refusals" – persons who originally refused to participate but were persuaded to participate.
- We realized that manipulation and interpretation of the data would have to be conducted with discretion. The results were not intended to be representative of the entire sample of the FAMEX respondents and no reliability tests were conducted to measure the significance of the results.

3. METHODOLOGY

The feedback survey was conducted on a test basis in two of Statistics Canada's regional offices: the Sturgeon Falls Regional Office covering northern and eastern Ontario; and the Pacific Regional Office covers British Columbia and Whitehorse. The 1996 FAMEX survey sample sizes were 1,304 for Sturgeon Falls and 2,528 for the Pacific Regional Offices.

The feedback questionnaire consisted of a one page questionnaire. It contained both "yes/no" check circles as well as space for write-in comments. The "yes/no" check questions asked about the introductory materials, the interviewer, and the use of a diary to replace the interviewer conducted survey. The "write-in" questions asked about respondent perception of the purpose of the survey, motivation for survey participation, and suggestions for, or comments about, the survey.

Respondents who completed or partially completed the Family Expenditure Survey were given the feedback questionnaire. Interviewers explained that it was an opportunity to provide information that would help to improve the Family Expenditure Survey. Completing the feedback questionnaire was voluntary and postage-paid envelopes were provided.

Results were analyzed at the aggregate level and by three household size and composition categories. The household size and composition categories are: one person households; households of two or more adults *without* children under 18; and households of one or more adults *with* children under 18.

4. RESULTS OF FEEDBACK SURVEY

4.1 Response Rate

The overall response rate to the feedback questionnaire was 21%. The response rates for individual questions at the aggregate level varied from a high of 96% for question 1 (Did you read the introductory letter?) to a low of 65% for question 8 which asks for comments and suggestions.

Usually, questions that required a written response instead of a "yes/no" answer have a lower response rate. This was not the case, however, for question 6 (Why did you participate?) which had a response rate of 94%. Possibly respondents wanted to take the opportunity to tell us they only participated in the FAMEX survey because it was mandatory. See the results for this question below.

4.2 Categorical Questions (questions answered with a "yes/no")

Four of the eight feedback survey questions needed categorical responses in that they simply required a "yes/no" answer. They covered survey introductory materials (letter, brochure), respondent perception of the interviewer, and respondent receptiveness to using a diary to record expenditures rather than the current method of interviewer-conducted survey.

4.2.1 Introductory Materials

For the 1996 FAMEX survey an introductory letter and brochure were designed to identify the survey authority, to explain the reason for the survey and to encourage survey participation. In the feedback questionnaire, Question 1, asks whether the letter was read and, if "yes", whether it was helpful. Question 2 is similar but concerns the brochure.

Of those that responded to the first two questions, 86% read the letter, and 77% read the brochure. Slightly over three quarters of those who answered the questions indicated that the material was helpful.

From the point of view of household composition, households with no children (both one person and those with more than one person) were more likely to find the introductory materials helpful.

4.2.2 Linking Respondents' Reaction to the Introductory Materials to Their Understanding of the Purpose of the FAMEX Survey

In order to gain an insight into the effectiveness of the introductory materials, we looked at the link between questions 1 and 2 (helpfulness of introductory materials) and the answers to question 5 (Why do you think we collect information on family spending?) for three groups of respondents:

1. those who read the letter and brochure and found them helpful;
2. those who read the letter and brochure and did not find them helpful;
3. those who did not read the letter and the brochure.

Of those who read the letter and brochure and found them helpful, almost two-thirds gave answers to question 5 that were correct answers *i.e.*, answers which showed they understood the reasons in the letter and brochure. Only slightly more than one-third of those who read the letter and brochure but did not find them helpful gave a correct answer to Question 5.

Most interestingly, over half of those who did not even read the brochure or the letter still managed to answer question 5 correctly.

These results imply two things:

- respondents that answered "no" to reading the letter and brochure may in fact have read them and did not remember; and/or
- the interviewer was able to relay the purpose of the survey effectively.

4.2.3 The Interviewer

Question 3 asked whether the respondent thought that the interviewer was professional, knowledgeable and helpful. The response rates for all three parts of this question were high and the "yes" answer was overwhelmingly chosen (98%, 97%, and 97% respectively). Results are similar for the three household composition types. Respondents obviously hold the interviewers in high regard and the interviewer-respondent relationship is definitely positive.

4.2.4 Alternative Data Collection Method (response rate 95%)

As stated earlier, the FAMEX programme is looking for ways to reduce respondent burden. Question 6, which asked about using a diary to collect expenditure information instead of the current interviewer assisted survey, was included to test receptiveness to this idea. Results were evenly split on this question. However, households with children were at least 10% more likely to prefer the diary method than one person households.

4.3 Qualitative Questions

Three of the feedback survey questions were "qualitative" in that they required respondents to write in their opinions or suggestions. Many respondents gave more than one comment or suggestion.

4.3.1 Why STC Collects Information on Family Spending (response rate = 85%)

Almost two-thirds of respondents who answered this question gave what could be considered to be correct answers. Answers were considered correct if the respondent made reference to any one of the following: calculating the Consumer Price Index, inflation, cost of living, pension or wage adjustments, or determining spending habits of families. Over a quarter of respondents gave "negative" reasons such as "no idea", "waste of money", "nosy government", "to raise taxes", and "to make more cutbacks". Analysis by household size and composition reveals some differences from the overall average. Households with two or more adults and no children are more likely to provide correct answers to this question – 71% answered correctly compared to the survey average of 61%. One person households were more likely to provide negative reasons (32% compared to the survey average of 27%).

4.3.2 Why the Respondent Participated in FAMEX (response rate = 94%)

Sixty percent of respondents who answered this question indicated that they participated in the FAMEX survey because it was mandatory. In addition, one third of those who answered this question indicated they participated because:

- they thought the survey was important and they wanted to help compile statistics; or
- they understood they had been selected and represented others; or
- they felt it was their duty as a citizen.

There is not much variation for this question by household size and composition.

4.3.3 Comments or Suggestions for Improvement (response rate = 65%)

The last question on the feedback survey asked for comments or suggestions to improve the FAMEX survey. Understandably, the most often received suggestion was to shorten the questionnaire (offered by 31% of those who responded to this question). Respondents were also concerned about the accuracy of the answers they gave due to the difficulty of recalling their expenditures over a one year recall period (19%) and wanted more notice of when the survey would take place. In fact, if all the comments that covered various aspects of "wanting more notice" are grouped, then this concern becomes the most often cited (32%).

Households with children under 18 feel the burden of the FAMEX survey most acutely. They are more likely to want more notice, want the survey to be shorter, would prefer to complete the questionnaire themselves and want a "thank-you"/incentive. They are also the most concerned about the accuracy of their responses.

One person households are less concerned about the length of the FAMEX interview but would like to be asked to prepare financial information ahead of time. They are the most likely of the three household types to suggest dropping the "mandatory" designation for the FAMEX survey.

5. UNFORESEEN USES OF THE RAS

There were three unforeseen uses of the feedback questionnaire:

- The opportunity to complete the feedback survey was used to persuade difficult respondents to participate in the FAMEX survey. The chance to vent their frustrations or to make suggestions for survey improvements directly to survey managers in Ottawa seems to be appealing to respondents.
- Early results from the feedback questionnaires were passed to the interviewers before they completed their caseload enabling them to improve their performance.
- Early results were also used to suggest topics or to fine tune existing topics discussed in focus groups exploring the respondent-interviewer relationship for the FAMEX survey. For example, concern about accuracy had not been anticipated as a focus group topic until it started to appear on feedback questionnaires.

6. CONCLUSION

As a means of exploring respondent reaction to the FAMEX survey at Statistics Canada, the response analysis survey (feedback questionnaire) has been very successful. Useful information was gathered on respondents' motivation for participating, their perception of the purpose of the survey and the uses of the data, and their reaction to the way this information was presented to them. This information will be used to work towards improved respondent relations and, hopefully an improved response rate.

The feedback survey highlighted respondent concerns with the length of the questionnaire and the recall time – especially for households with children under 18 years of age. It has also uncovered respondents' concern for the accuracy of their answers, their desire to know more about the survey, and their positive relationship with Statistics Canada interviewers. Statistics Canada is currently redesigning the FAMEX survey methodology and interviewer training to capitalize on these insights.

The use of response analysis surveys provides a relatively simple technique for developing or evaluating the data collection portion of the survey process and should be considered as a tool for the enhancement of respondent relations for complex surveys with high response burden.

REFERENCES

Bercini, D.H. (1991). Techniques for evaluating the questionnaire draft. *Statistical Policy Working Paper 20*; Seminar on Quality of Federal Data, Federal Committee on Statistical Methodology, Statistical Policy Office, Office of Information and Regulatory Affairs, 340-348.

Goldenberg, K.N., Butani, S.J., and Phipps, P.A. (1993). Response analysis surveys for assessing response errors in establishment surveys. *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, 290-299.

Phipps, P.A., Butani, S.J., and Chun, Y.I. (1995). Research on establishment-survey questionnaire design. *Journal of Business & Economic Statistics*, 13, (3), 337-346.

Utter. C.M. (1983). Response analysis surveys. *Statistical Policy Working Paper 10: Approaches to Developing Questionnaires*, (Ed.) Theresa J. De Maio, United States Bureau of the Census, 151-158.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

INTERVIEW AND PRESENCE AT HOME

J.L. Madre and J. Armoogum[1]

ABSTRACT

In surveys, whether they are conducted face to face or by telephone, the main difficulty in avoiding non-response and sampling bias is in reaching people at home at a time when they are sufficiently available to respond. Trip surveys are a good example of this problem, since:

- such surveys are a difficult case, in that what is being observed is mobility behaviour, and therefore information is sought regarding events that take place when people are not there to answer (giving rise to problems of recall, *etc.*);
- but once controlled, the information collected in a mobility survey (or a time use survey) can shed light on the problem of reaching respondents at home, for the purposes of any type of survey.

Drawing on the last two INSEE-INRETS national transportation surveys, we will distinguish three time horizons: the year (on the basis of long-distance trips), the week and the day (on the basis of daily mobility).

KEY WORDS: Collection; Non-response; Telephone; Personal interview.

1. INTRODUCTION

The 1993-94 transportation and communications survey produced a couple of oddities:

- In response to a question about travel on a weekday (the day before the survey), the proportion of trips made for the purpose of commuting to work was highest in December and July (21.4% and 20.8% respectively, compared with 15.5% in May).
- In the interview on long-distance travel in the previous three months, there was naturallly a substantial recall effect (less travel for the first month than for the month immediately preceding the survey), but there was also a lull for the two weeks before the interview.

We will show that these strange results are clearly due to the availability of households when contacted at their principal residence for a field interview, and we will illustrate the relevance of other examples (self-administered vehicle diary, long-term absence, *etc.*) by relating them as clearly as possible to the survey instrument used. We will also extract from the data on travel (*i.e.*, absence from home) useful information for other surveys in which personal or telephone contact requires that the respondent be at home. Transportation is not the only subject on which the respondent's presence at home can limit the scope of response, and the structure of the sample can be seriously distorted if the problem is ignored.

2. MONTHLY AND WEEKLY PATTERNS: LONG-DISTANCE TRAVEL

2.1 The Survey Instrument

The Transportation and Communications Survey is a massive project: 14,200 private households (collectives were excluded) were interviewed for a total of 1.5 hours on average in two visits at least one week apart. Only persons aged 6 or over were asked to describe their travel. To capture seasonal fluctuations, which are substantial in travel behaviour, the interviews were spread throughout the year (from May 1993 to April 1994, excluding the first three weeks of August). During the first visit, the interviewer asked one household member to describe his or her long-distance travel and provided the person with a memo-pad to help him/her prepare for the questions he/she would be asked during the second visit.

Departure and return dates were noted for all trips completed in the previous three months. There was very little non-response; only about 10 incomplete descriptions had to be excluded from the data. For the purposes of analysis, long-distance travel is traditionally defined as a trip of more than 80 km from one's home. The definition of tourism, which includes spending nights at locations less than 80 km from one's principal residence, would have shed more light on our subject.

Although the distribution of departure dates should be uniform, since seasonal variation was eliminated by conducting interviews throughout the year, a number of phenomena interfere:

- a recall effect (trips in the more distant past are forgotten),
- presence at home (a lull in the two weeks preceding the survey was due to respondents' postponement of the second interview),
- few date errors as far as we can tell (just a slight increase for the first week in the reference period).

2.2 Measuring Presence at Home using Travel Data

To each of the 90 days preceding the survey date, we assigned a probability of the respondent being at home: 0.5

[1] Jean-Loup Madre and Jimmy Armoogum, Institut National de Recherche sur les Transports et leur Sécurité, 2 av. du Général Malleret-Joinville, 94114 Arcueil Cédex, France; e-mail: jean-loup.madre@inrets.fr.

if it was a day of departure or return, 0 if the person was away on a trip, and 1 otherwise. Despite recall effects (short trips forgotten), the distribution of presence at home is very uniform for the beginning of the observation period (6-7% of respondents were away). However, the rates for the two weeks preceding the survey are lower: 1.5% the day before the survey and 3% for the rest of the last week. To overcome this survey effect, we computed the rates of presence at home for the period between the 11th and 89th days (inclusive) preceding the survey.

Starting from the interview date, we computed the average proportions of people away for the part of the sample in the coverage area just defined. If we look at French statutory holidays, for example, at least 24% of people living in France were away from home on August 15, 1993, 15% on July 14, 12% on Christmas Day, 8% on May Day, and 10% on New Year's Day, 1994. By contrast, roughly 99% were at home in the first week of December, for example. To summarize all these specific observations, we computed absence rates for each month, each day of the week, and three special periods (weekends, the week between Christmas and New Year's, and July-August).

We found that there was a much greater range within a year (3% of people were away in January and November; 20% in August) than within a week (6% on Monday through Thursday; 7.4% on Saturday). The first oddity described in the introduction to this paper is clearly due to the fact that 12% of working people are away in July and are therefore not available to answer questions about their daily travel, and this artificially inflates the proportion of people who, not being on vacation, commute to work.

2.3 Behaviour by Occupation and Social Group

The monthly and weekly behaviour patterns of retired people were quite similar to those of people in the same occupation who were still working. Absence rates clearly increased with income level, and for that reason, we grouped business leaders with senior managers, and service workers with unskilled labourers. Farmers stayed at home the most, though their absences have become substantially longer since the early 1980s. Each occupation had its own particular patterns: weekends shifted to Monday for retailers and artisans (low peak absence rate; on Sunday rather than Saturday), lowest absence rate for clergy on Sunday, vacations concentrated in August (plants closed) for labourers (especially skilled workers in large industry).

People who were away for extended periods and therefore could not be interviewed (1% of the population, mostly students and military personnel) accounted for at least 5% of the distances traveled in long-distance trips.

3. ABSENCE FROM HOME ON WEEKDAYS

3.1 Comparing Different Instruments

In the previous Transportation Survey (1981-82), daily travel was recorded in a seven-day travel diary. In the most recent survey (1993-94), respondents were interviewed about their travel the previous day (a weekday) and the previous weekend. For travel by car, we were able to compare those data with data collected using a seven-day vehicle diary. The interviewed respondents reported 9% more trips of over 45 km, and 10 departures for 9 foreign destinations; in the diaries, the two flows balance.

Once again these differences are due to the presence-at-home problem. To record events in a seven-day diary, one has to be at home when the canvasser drops it off and again when he/she picks it up a week later. Interviewing a respondent about the previous day's and previous weekend's travel requires just one contact and can capture data on returns from long trips that elude seven-day diaries. There is of course an imbalance between departures and returns (1/9 as many departures on vacation as returns, and half as many to a secondary residence), but interviews capture a substantial portion of trips farther than 80 km from home, most of which are not recorded by diaries. Because of these coverage differences, we limited our comparison of daily travel data provided by the 1981-82 travel diary and the 1993-94 interview to trips within an 80-km radius of home.

Finally, because of recall difficulties, travel by foot was not recorded on weekends. We also found, by matching the data against vehicle diaries, that 30% of trips of less than 2 km were forgotten. Hence we do not have a complete record of weekend travel. Although an analysis of travel on Saturday, when people exhibit specific patterns and some segments of the population are easier to reach, would have been very interesting, we had to confine our study to weekdays.

3.2 Contacting People at Home at Different Hours of the Day

Of the people whom interviewers were able to contact at home, 95% were home between 1 am and 4 am the day before the interview; the percentage falls to 44% between 10 am and 11 am, rebounds to 62% at 1 pm, and drops back to 40% between 3 pm and 4 pm. On weekdays, French residents over age 5 spend an average of 6.5 hours per day away from home.

Free from the obligations of employment, retired people exhibit a very different pattern of behaviour from employed people. This is in contrast to the findings for long-distance travel. Retired people were away from home the least amount of time (2.75 hours per weekday on average). At the other extreme, senior managers, professionals and business leaders were away from home for 10 hours a day.

To determine the main factors affecting presence at home, we applied a logit model to the proportion of people away from home at 3 pm. We found that sedentariness:

- increases with age: in particular, 88% of people aged 6 to 15 were at school at 3 pm, but their periods of absence from home were concentrated, totaling 7.25 hours a day, one hour less than the average for people aged 16 to 25;
- increases as population density shrinks (7 hours a day away from home in the inner suburbs, compared with 5 hours in predominantly agricultural areas); the fact that sedentariness is stronger in downtown areas than in the suburbs has more to do with the population structure

than with the unique characteristics of downtown life, "all other things being equal";

- people living in smaller communities are more inclined to stay home (5.75 hours a day away from home) while residents of the Paris metropolitan area are less so (7.5 hours a day);
- part-time workers spend 2.5 more hours at home than full-time workers, while unemployed people are at home 5.5 more hours;
- finally, married women are more likely to be at home, but the key factor is their marital status rather than their gender.

With all these factors, we were able to reconstruct 74% of the matching data, and we were even able to push it up to 79% by adding some less important variables. Finally, it is worth noting the substantial differences between the structure of the overall population and the structure of the population at home at 3 pm. Even though it is an extreme example, it illustrates the value of tailoring canvassing times to the habits of the target population and trying at different hours to reach respondents of telephone surveys.

4. CONCLUSION

Catching people at home is a rather widespread problem for household surveys, and it is especially serious when the subject is travel. Daily close-to-home travel and long-distance travel can be measured by proven methods – asking respondents to recall events, or having them complete a travel diary or a time use diary – but we are still unable to capture short trips in the vicinity of a temporary residence (while on vacation, during a business trip). Nevertheless, questions about the previous day's and previous week's travel provide the least incomplete overview of this type of travel, since it requires only one contact at the respondent's home.

The interviews indicated that, in the early part of a weekday, 3% of respondents were not at home, but were within 80 km of their homes. Adding the 6% who were more than 80 km away (and therefore covered by the part on long-distance travel) and the 1% who were away for an extended period (and could not be interviewed), we found that about 10% of the population were not at home in the early hours of a weekday morning; roughly 15% were away early Sunday morning. Consequently, in our analysis of presence at home at different times of the day, we can subtract 5% from the presence rates to allow for people whom the interviewer was unable to contact because they were away for at least two days (reference day and survey day).

This paper also raises a more general issue: the statistical unit. Should we be satisfied, as in the Transportation Survey, with selecting respondents at random and letting the interviewers manage the timing of contacts during a six-week survey campaign? Or is it better to predetermine the periods for which we want to observe the travel behaviour of the randomly chosen individuals? If we want the latter, we have to find a flexible data collection method that does not require the respondent to be home (*e.g.*, a self-administered questionnaire designed to achieve a high rate of return). If we cannot find such a method, the non-responses will be high, correlated with the travel behaviours we are attempting to measure, and thus adjusting for it will be difficult.

ACKNOWLEDGEMENTS

The research described in this paper was conducted with the assistance of M. COMYNE, a PhD student at INRETS. Discussions with L. LEBART (Ecole Nationale Supérieure des Télécommunications), F. DUPONT and M. CHAVENON (Centre d'Etude des Supports de Publicité) were very helpful. The project was funded by the government of France (Direction de la Recherche et des Affaires Scientifiques et Technologiques, Ministère des Transports) and by the European Union (DG VII 4th FPRD) under its Methods for European Surveys of Travel Behaviour (MEST) program, which includes the Institut für Strassenbau und Verkehrsplanung, Leopold-Franzens-Universität Innsbruck, Statistics Netherlands, Bro Herry (Wien), University of London Centre for Transport Studies, Imperial College (London), Deutsche Versuchsanstalt für Luftund Raumfahrt (Köln), INRETS (Arcueil), Transportes, Inovao e Sistemas (Lisbon), TOI (Oslo), Transport Studies Group Facultés Universitaires Notre Dame de la Paix (Namur), Socialdata (München), Statistics Sweden, TU Delft. The conclusions reached by the authors of this paper do not necessarily reflect the views of the European Commission or the Consortium de chercheurs.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A STUDY OF NON-RESPONSE ON THE UK FAMILY EXPENDITURE SURVEY

J. Martin[1], K. Foster, J. Hansbro and C. Ash

ABSTRACT

Since April 1995 interviewers working on the UK Family Expenditure Survey (FES) have used a short questionnaire to collect systematic information about non-responding households. Limited information about household composition can be collected for most non-responding households but other information is less complete. Information about the characteristics of non-respondents show broadly similar results to those from an earlier study which linked sampled addresses in 1991 to Census records for the same addresses in order to examine non-response bias. However, there were some differences which might reflect changes over time and a decrease in response rates. For a month, interviewers completed a more detailed questionnaire for both responding and refusing households, recording details of the interaction with potential respondents and the exact outcome of each call. This allowed comparison of reluctant and co-operative respondents and investigation of interviewer strategies for overcoming reluctance.

KEY WORDS: Non-response.

1. INTRODUCTION

In the UK information on the characteristics of non-respondents for major government surveys is available at ten-yearly intervals from studies which link sampled addresses to their decennial Census records. This allows non-response bias to be measured for any census variable (Foster 1997). It gives an accurate picture of non-response bias for a particular survey at a particular point in time. It does not, however, tell us about bias in survey variables which were not included on the Census which are usually the most important survey estimates.

For operational reasons, the matched datasets are not available until a few years after a Census and so estimates will be considerably out of date before new data are available. Moreover, they are based on particular samples and thus subject to sampling error. In addition, both response rates and bias due to non-response may change over time so it may not be appropriate to use the information directly for weighting to compensate for non-response.

We therefore investigated what information it might be possible to collect routinely about the characteristics of non-responding households with sufficient reliability either to weight the sample or to make adjustments to weights derived from the census studies in-between Census years.

Previous research has shown that interviewers are able to collect a certain amount of information about the characteristics of non-respondents, particularly when contact has been made with the household (*e.g.*, Groves and Couper 1993; Lynn 1996; Campanelli, Sturgis and Purdon 1997). Information that can be obtained by observation of the accommodation and area is clearly easiest for interviewers to obtain for all addresses. But even for refusing households, it is often possible to obtain some information about household members either by direct questioning or by observation.

We also report on a small study in which more subjective information was collected about the interaction between interviewers and potential respondents in both co-operating and refusing households, looking particularly at whether those who co-operate reluctantly are more like refusers or those who co-operate readily, and what implications this may have for both non-response bias and non-response reduction.

2. BACKGROUND

The study described was carried out on the Family Expenditure Survey (FES), a continuous survey of some 7,000 households per annum which provides data on household income and expenditure in the UK. Information is collected by means of a lengthy interview on expenditure and income with all adult members of the household, followed by a two week diary of all items of expenditure made by each person. A very strict definition of response is used: a household is coded as responding only if *all* adults in the household co-operate with both the interview and diary. The survey has been running since 1957 and for a long time achieved a response rate of around 70%. However, this had fallen to 65% by 1995/6, the year to which this study relates, despite efforts to maximise response through fieldwork strategies such as incentive payments and interviewer training.

The FES has a low non-contact rate – less than 3%. The largest category of non-response (27% of those eligible) is

[1] Jean Martin, Director of Survey Methodology, Social Survey Division, Office for National Statistics, 1 Drummond Gate, London SW1V 2QQ, UK; e-mail: jean.martin@ons.gov.uk.

those who refuse before the interview. Losses due to households failing to complete either the interview or the diary are small because interviewers are instructed to attempt to get all household members to agree to participate fully before starting any interviews. Respondents are paid an incentive of £10 ($16 US) each for completing the diary provided that *all* household members complete their diaries.

The high refusal rate (32%) relative to other household surveys reflects the high respondent burden and the stringent response rules. Foster (1997) shows that differences in response rates for five major surveys reflect both field operational procedures and respondent burden, diary surveys in general having lower response than interview only surveys, and surveys requiring participation from all adults having lower response than those that allow a single respondent.

Because response rates on the FES have been falling in recent years there has been particular interest in understanding more about the nature of the non-response with a view to trying to reverse this trend and to improve methods of compensating for non-response.

3. COMPLETENESS OF DATA ON NON-RESPONDING HOUSEHOLDS

Following a small pilot study (Cheesbrough, 1993) a non-respondent questionnaire (NRQ) was introduced on the full FES sample from April 1995 to be used by interviewers for all non-contacts and for refusals before or during the interview. At refusing households, interviewers tried to obtain the information by asking the questions directly of a member of the household. Failing this they could try to obtain information from a neighbour or code the items from their own observation or impressions. The NRQ records which was the main source of the information obtained.

Although the NRQ was returned for almost all eligible households (99.95%), the completeness of the data on each question varied considerably (Table 1). Not surprisingly the type of accommodation, which can be coded by observation, was available for all cases. Questions collecting simple factual information had the smallest proportion of missing answers. Household composition information – number of adults and children in the household and sex and age of the head of household in broad bands – was available for most refusing households: no more than 14% of these items were missing. However, such information was more likely to be missing from the small proportion of non-contacts.

By far the largest proportion of missing data was for the question on household income band (85%). This is clearly a sensitive question to ask of people who have already refused to take part in the survey and it cannot be recorded by observation. In addition, a number of questions were only available for around one half or three fifths of non-responding households. These included tenure (37% missing), availability of a car or van (44% missing) and ownership of a telephone (36% missing); all of which may be difficult for an interviewer to complete if he or she has not been able to speak to a member of the household.

Table 1
Item non-response on the FES non-response questionnaire

Variable	Refusals	Non-contact	Total non-responders
	percentage of missing answers		
Type of accommodation	0	0	0
Number of adults	5	39	9
Sex of head of household	7	44	11
Number of children	11	42	14
Age of head of household (including banded answers)	10	58	14
Marital status of head of household	18	67	23
Employment status of hoh	29	66	32
Telephone ownership	32	72	36
Tenure	35	59	37
Age of head of household	28	58	39
Availability of car/van	41	78	44
Household income	84	94	85
Base	2796	281	3077

Overall, a household member was the main source of information in just over half (53%) of cases, 42% of cases were coded according to the interviewers' observation or impressions, and a neighbour provided information in only 5% of cases. However, neighbours were the source for 41% of non-contacts, suggesting that interviewers can obtain a limited amount of information in this way. A member of the household was the main source of information for the majority of refusals, particularly where refusal occurred during the interview (86%), rather than before the interview (57%).

The proportion of missing items varied according to how the interviewer collected the information although the general pattem was the same for each category. With the exception of type of accommodation, the proportion of cases with missing data was lowest where the main source of information was a household member.

3.1 Discussion

The level of missing data and the quality of the collected data as indicated by its main source are of greater importance if the data are to be used for weighting rather than simply to monitor the characteristics of non-responding households. In order to derive weights, information on the characteristics used to define weighting classes is needed for all, or at least most, responding and non-responding households. The results suggest that only broad details of household composition would be suitable for this purpose. Given that population statistics for the age and sex composition of the population are available as control totals, number of adults in the household and possibly the presence, if not the number, of children would seem to be the most useful additional variables.

4. CHARACTERISTICS OF NON-RESPONDING AND RESPONDING HOUSEHOLDS

For non-responding households at which at least some information was collected we can compare their characteristics with those of responding households. Although we can do this separately for refusing and non-contact households, on the FES the refusals so outnumber the non-contacts that the picture of non-responding households as a whole will be largely influenced by the refusers.

Full details of the comparison of non-responding and responding households are given in Hansbro and Foster (1997). Table 2 summarises some of the more important differences found. It needs to be recalled that the level of missing data increases down the table and is very much higher for non-contacts than refusals.

Non-contacts differ significantly from responders on all the items shown but there is less evidence of differences between refusals and responders so the overall bias is not so serious. Refusals are higher among older people living alone and those with no children, and also somewhat higher for households with three or more adults. It should also be noted that people sharing accommodation but keeping their housekeeping arrangements separate are counted as separate households and could respond independently of one another.

Most of the variables associated with non-response bias were also included in the 1991 Census so we can compare the results with those from the linked datasets described above – shown in the last three columns of Table 2. These figures show very similar patterns. The figures for non-contacts are different in level but both are based on quite small samples. The census linked results do not show more refusals among elderly single person households but show greater differences for households with three or more adults.

4.1 Discussion

The characteristics of non-respondents collected on the NRQ were similar to the characteristics of non-respondents reported in the Census-linked study although a few differences were evident. It is clear that non-contacts differ markedly from responders in many respects. However, because on the FES they form a small proportion of total non-response they do not affect the characteristics of non-responders in general very much. The nature of these studies does not allow us to determine whether the differences are due to genuine changes over time; other factors may also be relevant such as the drop in response rates over the period, sample differences or unreliability of the NRQ data.

5. CHARACTERISTICS OF RELUCTANT RESPONDERS COMPARED WITH CO-OPERATIVE RESPONDERS AND REFUSERS

In addition to completing the non-respondent questionnaire, interviewers were asked to provide more detailed information during October 1996 about each household with whom contact was made, whether the outcome was an interview or a refusal, by completing an interviewer contact questionnaire (ICQ). The questionnaires for responding and refusing households were slightly different although many of the same questions appeared on both questionnaires.

The main aim was to look at how interviewers dealt with reluctance and refusal and which strategies seemed to be most successful in persuading people to overcome their reluctance. The information collected also allows us to look at the characteristics of reluctant households and compare them with readily co-operating and refusing households.

Interviewers completed ICQs for 80% (413) of responding households and for 71% (225) of refusing households. In addition to missing forms, there were also missing answers for some of the questions. For responding households, interviewers classified the respondents as 'co-operative' (292) or 'reluctant' (109). The latter were households where the interviewer judged that he/she had had to make a particular effort to persuade one or more household members to take part in the FES.

Table 2
Household characteristics associated with non-response

	Non-respondent questionnaire			Census linked dataset		
	Refusals	Non-Contact	Responders	Refusals	Non-Contact	Responders
	%	%	%	%	%	%
1 person 16-29	3	18	3	6	41	10
1 person 60 +	20	23	15	17	20	17
3 or more adults	17	2	14	28	8	17
No children	80	93	68	78	87	71
Unemployed	4	8	5	5	14	5
Apartment	20	53	18	16	52	17
Private renter	6	16	10	7	14	8
No telephone	9	38	8	NA	NA	NA
Base	*2703*	*159*	*6650*	*1277*	*71*	*3359*

Comparing the characteristics of refusing and responding households to those described above on the full 1995/6 sample showed broadly similar results so there is no indication that there were serious biases in this small sample. The information from the ICQ enables us to compare the characteristics of co-operative and reluctant responders with one another and with the refusing households.

Reluctant responders were less likely than either co-operating households or refusing households to contain three or more adults and slightly more likely to contain only one adult. Although for some characteristics there was no clear distinction between co-operating, reluctant and refusing households, in other respects reluctant households were more like refusing households than co-operating households. In particular they were more likely than co-operative households to have a head of household who was aged 70 or over: 30% of reluctant households had a head of household in this age-group compared with 15% of co-operative households. It seems that interviewers are particularly likely to overcome the reluctance of elderly single people to respond.

Table 3
Reluctant and co-operative respondents compared with refusers

Household type	Refusal	Reluctant responder	Co-operative responder
	%	%	%
1 adult	33	39	34
3 or more adults	14	5	11
HRP 16 - 29	16	9	10
HRP 70+	23	30	15
Unemployed	7	14	8
Retired	36	37	25
Base	*217*	*109*	*292*

6. STRATEGIES TO OVERCOME RELUCTANCE

Interviewers who had carried out an interview at a reluctant household were asked what strategies they had used to overcome the respondents' reluctance. In 42% of cases the interviewers had backed off and then called again later; in 32% of cases the interviewer had explained the purpose of the survey. The other two strategies commonly used were to explain to the respondent that it was important for everyone to have their say (13% of cases) and to reassure the respondent about confidentiality (10% of cases). Reminding people about the £10 payment was used in 9% of cases. Looking at the details of the interactions recorded by interviewers revealed that they did try to adapt what they said to attempt to persuade refusers to take part.

7. CONCLUSION

This study has established that collecting a limited amount of information about the characteristics of non-responding households is feasible and can potentially provide useful information about non-response bias which may be of use in weighting for non-response. It remains to be tested whether more complete data could be collected if we sought information about a smaller number of items. In addition to the household composition variables it would be very useful to know housing tenure. We plan to do further work to see if we can improve success in establishing this.

We have also been evaluating calibration methods for producing household weights for this survey which correct some of the biases in household composition. We need to determine whether it would be worth adding preliminary weights based on this sort of information to population totals.

Although further information about the interaction between interviewers and respondents was carried out only on a small scale, it was possible to distinguish between co-operative and reluctant responders and to show that in some, but not all, respects the latter were more similar to refusers than to co-operative responders, which indicates the importance of persuading such people to take part in the survey from the point of view of reducing non-response bias. We intend adding some extra questions to the main NRQ in future to identify both reluctant respondents and circumstantial refusers who might on another occasion or with a different survey have agreed to cooperate and to learn more about what interviewers can do to overcome the reluctance of the latter.

This paper is a shortened version of the full paper which was prepared for the Statistics Canada Methodology Symposium 97. Copies of the latter are available from the contact author on request.

REFERENCES

Campanelli, P., Sturgis, P., and Purdon, S. (1997). *Can you hear me knocking?: An investigation into the impact of interviewers on survey response rates.* London: SCPR.

Cheesbrough, S. (1993). Characteristics of non-responding households on the Family Expenditure Survey. *Survey Methodology Bulletin*, 33, 12-18.

Foster, K. (1996). A comparison of census characteristics of respondents and non-respondents to the 1991 Family Expenditure Survey. *Survey Methodology Bulletin*, 38, 9-17.

Foster, K. (1997). Evaluating non-response on household surveys: report of a study linked to the 1991 Census. *GSS Methodology Series* (Forthcoming).

Groves, R.M., and Couper, M.P. (1993). Unit non-response in demographic surveys. *Proceedings of the Bureau of the Census 1993 Annual Research Conference.* Washington: Bureau of the Census, 593-619.

Hansbro, J., and Foster, K. (1997). Collecting data on non-respondents to the Family Expenditure Survey. *Survey Methodology Bulletin*, 41, 27-36.

Lynn, P. (1996). *Who responds to the British Crime Survey*? Paper presented at the International Social Science Methodology conference. Essex University, 1-4 July 1996.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

DEALING WITH NON-RESPONSE WITH A POLITICAL COMPONENT

N. Moon[1]

ABSTRACT

Differential refusal was a significant factor in causing the error in both the pre-election polls and the exit polls at the British General Election of 1992, with Conservative voters more likely to refuse. It would be difficult to get a lower level of refusals on an exit poll than the current 16%, so NOP and the BBC attempted to develop of a model which would allow an estimate to be made of the voting behaviour of those who refused. Two approaches were tested: to find an alternative question which people would be prepared to answer and which would give a reasonable indication of their voting behaviour; and to get the interviewer to estimate something about the respondent which could be used in predicting their voting behaviour. Asking an extra question proved unsatisfactory, because most people who refused the voting question then refused the alternative question as well. The one interviewer estimate which was tried and found to be successful was getting interviewers to estimate the voting behaviour of their respondents before they initially approached them. Interviewers proved better at estimating voting intention than chance alone would suggest and across the series of experiments applying this corrective weighting always either made the result better, or at least made it no worse. This method was therefore used in the 1997 general election exit poll, where it was again found that the Conservatives were more likely to refuse to take part, and use of the weighting based on interviewer estimates considerably improved the exit poll estimate of the national share of the vote.

KEY WORDS: Exit poll; Refusal; Bias.

1. INTRODUCTION

In the 1992 general election the error on the BBC/NOP exit poll, while small, was outside sampling error, and this led to discussions between NOP and the BBC on how the error could be reduced in any future exit polls. Given the nature of an exit poll, there are only two real causes of error – sampling bias, and differential response rates. Detailed investigation of the figures showed that differential refusal rates among supporters of different parties was the only serious explanation of the error on the exit poll. This fits in with the conclusions of the Market Research Society committee of enquiry into the pre-election opinion polls, which gave considerable weight to differential refusal. It is worth noting at this point that both the direction and level of error were virtually the same on the two BBC/NOP exit polls – the marginals prediction poll and the nationally representative analysis poll – and that they were also the same on the two Harris/ITN exit polls.

Although the level of refusal on exit polls is fairly low at around 16%, it is high enough to cause significant errors if supporters of one party are more likely to refuse than those of another. It is unlikely one could reduce the already low level of refusal, so if there will always be a small but significant minority who will not say how they voted, the alternative approach is to try to estimate how the refusals did in fact vote.

Some work has been done on the pre-election polls on ways of estimating the vote of don't knows and refusals, though here we had the advantage that people who refused the voting question may well have answered a whole series of other questions. Because an exit poll is virtually a one-question questionnaire, those who refuse the voting question refuse the whole survey, and we have no information about them other than the interviewers' estimates of their sex and age. If it were possible to get some other information about them, it might be possible to reallocate the refusals in a sensible way between the parties. Given the circumstances of an exit poll an interview of any length is out of the question, and so it was decided that some experimentation would be carried out to try to discover a single question which those who refused the voting question would answer in reasonable numbers, and which would give a reasonable indication of how they had voted.

Such experimentation can only be carried out at an election, and thus the opportunities are limited. An initial, exploratory exercise was carried out in the Parliamentary by-election in Newbury, though the experimental programme proper did not begin until the by-election in Dudley West. This first stage was followed by separate experiments at the Scottish and English local government elections in 1995 (fortunately there was a gap between the two in which the results of the first could be assimilated before finalising the design for the second), and a final round of experiments at the English local election results of 1996.

The following section summarises the experiments very briefly, although a fuller version of the paper, including tables showing the success or otherwise of each experiment, is available on request from the author.

[1] Nick Moon, NOP Research Group, Ludgate House, 245 Blackfriars Road, London, SE1 9UL; e-mail: nickm@nopres.co.uk.

2. EXPERIMENTAL METHODOLOGY

The experiments can be split into two types - those which involve asking respondents an extra question, and those which involve interviewers estimating something about voter characteristics.

2.1 Extra Questions

2.1.1 Newspaper Readership

The first method tested was an additional question on newspaper readership. This was used in two Parliamentary by-elections in Newbury and Dudley West. In order to test the impact of the question itself, one half of respondents in Newbury were asked the question before being asked to complete the ballot form, while in the other half only those who refused to answer the ballot were asked the readership question. In Dudley the question was only asked before the ballot form was handed over, but this was one of three different questions each asked of a quarter of respondents (the remainder being a control group).

In each case the event the experiment did little more than broadly confirm our earlier suspicions about refusals. Because many people who refused the voting question also refused the question on readership, and because many of those who did answer did not read a national daily newspaper regularly, the base sizes were very small, especially in Dudley. All that we could conclude was that readers of Conservative-supporting newspapers were more likely to refuse than readers of other papers, suggesting again that Conservative voters were more likely to refuse. This was useful, but we needed something we could use actually to weight the raw data so we could come with a new, estimated, result.

The next test after Dudley was the Scottish local elections, and because the pattern of newspaper readership is so different in Scotland the question would not have worked, and so it was abandoned at this point.

2.1.2 Best Prime Minister

One variable which we knew from our experiments on opinion polls to be strongly correlated with vote was the question on who would make the best Prime Minister. This was included as one of the three alternative extra questions in Dudley.

It must be stressed again how low the sample sizes were for these split sample tests. For the question on approval rating of John Major, there were only 17 people who answered this question but did not complete the ballot. One should not really produce percentages based on a base as low as this, but it is difficult to make comparisons between refusals and the rest otherwise.

If the refusals were a random subset of the electorate, their rating of John Major would match that of all respondents, while if they are much more likely to be Conservatives the pattern of their answers should be close to that of Conservative voters. In fact, refusals were indeed slightly more Conservative than the whole sample, but their answers were much closer to the base all figures than to the Conservatives. Any weighting of the results based on this data would have had only a marginal effect.

After Dudley it was decided that split samples were impractical, and only one extra question would be asked. Since it had worked less well than the alternatives, the question on best prime minister was therefore abandoned.

2.1.3 Best Party for the Economy

Another question which we knew correlated well with voting was that of which party would be best at running the economy. This was the last of the three alternative questions used in Dudley, and was the only extra question asked in all subsequent experiments. As a discriminator this question is much better than the rating of John Major, since in Dudley 95% of Conservative voters thought the Conservatives were better at running the economy, and 98% of Labour voters thought Labour were better, with similar figures obtained in the other tests.

The only problem with the economy question is that it only offers a choice between the two main parties, and yet we need to use it to weight the votes of all the parties. In Dudley we proceeded as if all refusals were either Conservatives or Labour. While weighting in this way improved the Labour and Conservative forecasts, it made the Liberal Democrat figures worse.

In weighting all refusals according to the best party for the economy it was not sufficient simply to assume a complete correlation between voting and best party for the economy. Instead we decided to assume that those who answered "neither" were supporters of other parties. Generally this meant the Liberal Democrats, but the situation in Scotland was complicated by the fact that it was a four horse race if we include the Scottish Nationalists. To make up for this, those who refused the ballot and said 'neither' at the economy question were divided between the SNP and the Liberal Democrats in proportion to the overall numbers who voted for each of them (3:2) On this basis an adjusted exit poll showed a slight rise in Labour and Tory support and a slight drop in SNP support. Unfortunately it did not bring us any closer to the actual result.

The same approach was used in the two experiments at English local elections, apart from the need to allow for the SNP. In each case the weighted results were better than the unweighted, but the differences were small and the unweighted ones had been acceptably close to the actual result anyway.

2.2 Interviewer Assessments

In the 1992 general election interviewers had estimated the age, sex and class of all voters, so that demographic weighting could be applied to compensate for refusals. Age and sex had no effect, and class a minimal one, so class was included in the experiments but not age or sex.

2.2.1 Estimated Social Class

Interviewers estimated the social class of half the people approached in the English experiments, and all those approached in the Scottish experiment.

This was used in the weighting by applying the voting pattern of those in each class group who **do** complete the ballot to those in the same class group who do not. Thus of those C1s who did complete the ballot in Dudley, 65% said they voted Labour, and 23% Conservative. We then assume that of the 26 in class group C2 who refused the ballot, 65% – 17 people – were Labour voters, and 14% – 3 people – were Conservative voters. Following the same calculations for the other parties, and for the other class groups, produces an estimate of how the refusals might have voted, based on their social class.

In three of the four cases, weighting in this way made the result worse, while in the final test it made no difference.

2.2.2 Estimated Voting

The other estimate made by interviewers was of the actual voting behaviour of all selected respondents. They were asked to classify all people approached as either Conservative, Labour, Liberal Democrat or other. It must be stressed that this was done **before** the interviewer gave respondents the ballot form. For this measure, as well as being able to look at how refusals differed from respondents, it is also possible to measure the ability of the interviewers to "read" respondents in this way. The vast majority of people approached do fill in the ballot form, so for all of these we can compare the interviewer estimate with actual behaviour. If interviewers do not do better than pure chance, we would not use the estimate for weighting, while if they do significantly better we would use it.

In fact interviewers generally did do better than pure chance, and on the three occasions this test was used (it was not used in Scotland because of the complication of the SNP) it either made the poll results better or at least made them no worse.

Table 1 summarises the results of all the different experiments.

Table 1
Effect of each experiment

	Newbury	Dudley W	Scottish local 1995	English local 1995	English local 1996
newspaper readership	indicative only	indicative only	n/a	*made pole worse*	n/a
prime minister	n/a	made no difference	n/a	n/a	n/a
economy	n/a	**made poll better**	made no difference	**made poll better**	**made poll better**
estimated class	n/a	*made poll worse*	*made poll worse*	*made poll worse*	made no difference
estimated vote	n/a	**made poll better**	n/a	**made poll better**	made no difference

3. THE ACID TEST – THE 1997 GENERAL ELECTION

After the 1996 English local elections it was possible that there might be another Parliamentary by-election in which to carry out one further test but as this could not be guaranteed, it was decided that no further experimentation would take place. In any case, by this time it had become fairly clear what the workable alternatives were. Of the extra questions it was clear that the only one which had any merit was the economy one, which had generally had small but beneficial effect. However there were two significant concerns about using any extra question. The first was that having interviewers ask any kind of question would put off some people who would otherwise have completed a self-completion form.

The argument that the extra question may increase the refusal rate was a powerful one, but it was reinforced by another. The whole principal behind all this experimentation was that refusals are different from respondents, and that some means must be found reliably to estimate their behaviour. All the experiments involving an extra question showed that most of those who refused the voting question also refused the extra question. In weighting the refusals using the answers to the extra question we had to weight all the refusals using the answers just from those who did answer the extra question, thus assuming that those who refused the extra question were the same as those who answered it: when the whole approach was based on a belief that this was not so for the voting question.

Either of these arguments alone would have been powerful, but together they convinced us we should not use an extra question in the general election. This left the issue of whether to use interviewer estimation, and if so what. It was clear that estimating age and sex was of little value, and since it only added to the volume of data to be phoned back, they were not used. The experiments showed that weighting by estimated class was more likely to make matters worse than better, so the only remaining option was estimated vote.

The experimental results had been very encouraging, but there were potentially severe presentational difficulties. The whole team had visions of one of the tabloid papers running a story on how the BBC did its exit poll just by getting interviewers to guess, but it was eventually decided that because we had the ability to verify the method throughout the day, by comparing estimated vote with actual vote, estimated vote would be used to correct for refusals.

As a further check on the method a second, parallel, method known as ecological inference was used. This involved examining the level of refusals against the political complexion of each location. We already knew that if the level of refusals is much higher in Conservative wards than Labour ones, one could reasonably assume that refusals are more likely to be Conservative. The process of ecological inference takes this process a stage further, and aims by statistical analysis to make an actual estimate of the effect of refusals. It was agreed that weighting for refusals would only be used if both the weighting and the ecological inference produced similar figures.

Once again the interviewers proved significantly better at identifying voters than pure chance, showing the method was reliable enough to be used. Given the election was marked by a lot of switching by traditional Conservative voters to Labour it is not surprising that many whom the interviewers thought were Conservative in fact voted Labour. For Conservatives, the correlation between interviewer estimate and the question on voting in the previous election was much higher. Although estimation of Liberal Democrat voters was much less good than for the two main parties, this was not felt to be a major problem because the real issue over the refusals was how many of them were Conservative.

Table 2
Reliability of interviewer estimates

	Estimated vote			
	Conservative	Labour	Liberal Democrat	Other
Actual vote	%	%	%	%
Conservative	**46**	19	27	15
Labour	34	**64**	38	38
Liberal Democrat	16	12	30	17
Other	5	6	6	31

The unweighted data from the poll were again some way off the actual election result. Some of this was due to the sampling of constituencies, but it is clear that differential response was again at work. In the end the broadcast figure was manipulated by the interviewer estimates, by the ecological inference, and by an estimate based on claimed vote at the 1992 election. The net effect was to move the exit poll 2% closer to the true answer for each of the two main parties.

Table 3
The effect of weighting in the 1997 general election

	Raw figures	Weighted figures	Actual result
	%	%	%
Conservative	27	29	31
Labour	49	47	45
Liberal Democrat	17	17	17
Other	6	6	7

We would have no hesitation about using the technique in future exit polls. Because it can always be validated by internal data it has a built-in safety device, and because it is based on actual data rather than extrapolation from other questions, it will still work even if the political climate changes and Labour voters become more likely to refuse.

SESSION C-5

Handling Census Information

NEW GRAPHICAL TECHNIQUES FOR THE ANALYSIS OF CENSUS DATA

D. Desjardins[1]

ABSTRACT

This paper highlights a special "Industry Profile" plot that was developed by the author as part of our EDA implementation plan for the analysis of the U. S. Census Bureau's Economic Survey data. This multivariate plot provides analysts with a comprehensive overview of the companies covered by the Bureau's Economic surveys. This single plot profile covers most of the key variables on a survey form. This is accomplished by combining/displaying the inter-relationships of a number of these key variables (that are individually itemized on each of our survey forms), thus, in one graph, allowing our Analysts to quickly and easily see the unique characteristics of (and problems with) their data.

KEY WORDS: Exploratory Data Analysis (EDA); SAS/JMP; SAS/INSIGHT; Complex graphics; Industry profile plots; Leverage plots; Census survey data.

1. INTRODUCTION

Ongoing large-scale periodic Economic, Agricultural, and Demographic Census surveys at the U.S. Bureau of the Census produce massive floods of data. When it comes to the editing and analysis of this flood of data, many Analysts literally beg for faster, more effective techniques. Currently, in many areas of the Census Bureau, we still use traditional, ("blind", black box) CPU techniques that produce massive tabular printouts of items to edit. These edits are typically of limited sophistication – consisting of things like simple survey response cell checks (internal consistency), comparisons between ranked past-year to current-year differences (tolerance edits), and the use of a single (key) variable to impute missing values. Further, since working with these printouts is so slow, it allows our Analysts little time to investigate more than a few key items – much less the time required to understand the broader implications of our data and assuring that we maintain our highest standards of quality.

Today, we have an unprecedented opportunity to implement any number of special Graphical Exploratory Data Analysis (EDA) techniques for the analysis of this flood of data. There are two key factors that make this unprecedented opportunity possible. The first is the remarkable power of today's low cost personal computers – making possible graphical data analysis with unprecedented speed, superb flexibility, and extremely high interactivity. This is a change as revolutionary as was gunpowder to warfare. Today, we can generate more graphs in mere minutes than it would have taken months to generate just a few years ago. In addition, incredible new display powers are now available within these graphs – for instance, each of the points in the simultaneous display of a dozen different graphs can be highlighted (cross-linked to one another) by a technique called "brushing". The second factor is the use of proven off-the-shelf graphics software alongside the solid framework of EDA methodology that is already established. For instance, beginning with the now over 25 year old landmark book: EXPLORATORY DATA ANALYSIS, by the widely renowned Statistician, John Tukey, graphics-based EDA techniques have slowly, but steadily, found their way into standard data analysis practice. Today, numerous new EDA techniques are regularly extolled in many texts and technical journals. These techniques include powerful macro-editing procedures put forward by individuals such as Hidiroglou, Berthelot and Granquist, – as well as automated editing and imputation methodology by Fellegi and Holt. However, until now, specialized graphical EDA software application packages were typically required to be paired with this methodology (for instance, ARIES from the Bureau of Labor Statistics, and GEAQS from the Drug Enforcement Administration). This key roadblock has now been removed.

As part of our ongoing modernization program, the Bureau recently signed an approximate $2 million dollar per year contract with SAS corporation. This contract includes two of SAS's new, powerful, highly interactive EDA software packages – JMP and INSIGHT. These relatively easy to learn (point-and-click) general purpose packages stand ready to find suitable EDA applications. Thus, we now have a unique mix of powerful hardware, general purpose software, established methodology, and unprecedented need. Like a match to gunpowder, all that is additionally needed is a successful implementation plan.

[1] Dave Desjardins, U.S. Bureau of the Census, Washington, DC 20233, U.S.A.

This paper does not represent the position of the Bureau of the Census – all opinions are those of the author. In addition, to insure against the possible disclosure of confidential information, only fictitious data sets were used for the illustrations in this paper.

2. IMPLEMENTATION STRATEGY

"The value of any statistical methodology is directly related to how easy it is to understand and inversely to how hard it is to implement."

Anon.

Unfortunately, introducing, any form of change, much less the application a number of innovative Graphical EDA techniques into a bureaucratic organization filled with traditionally trained Statisticians can be a tough nut to crack. All too often, change is not viewed as an exciting challenge, but rather as a major headache. One would think that simply showing the advantages of EDA techniques would be sufficient to insure implementation. The Bureau is certainly not alone in this – one has only to scan the vast majority of technical papers in professional journals to see a virtual dearth of graphs – a fact that, no doubt, indicates a corresponding nation-wide avoidance of the use of EDA techniques. Because of this, we deemed the first step of an effective implementation strategy must have numerous sound/practical applications to go hand and hand with these powerful EDA techniques.

Another limiting factor is the incredible power of the EDA software package itself. Because SAS is so powerful and contains so few actual online examples/illustrations, it also has a very steep learning curve (it is viewed by many novice users as a "swamp"). Accordingly, our first step was to devise a way to help insure that the Analysts at the Bureau (who usually are casual users with only basic statistical expertise) were able to easily implement these new procedures. This first step entailed taking 10 key EDA techniques and condensing them into a set of point-and-click code – a "cookbook" of very powerful, easily generated, general purpose graphic techniques. Using this cookbook, even novice SAS users can generate any of these graphs in just a few minutes. The second step entailed insuring that these new EDA techniques were not only easily implemented, but understood. To satisfy this requirement, the author is also currently conducting an on-going series of introductory-level 40-hour EDA/Graphics courses for Bureau Subject Matter Specialists. This course is special because it is designed for non-statisticians, and requires a minimum of formal statistics. Instead, this course emphasizes an understanding of basic EDA graphs – as well how to apply these graphs in a step-by-step process of easy to understand, practical, data analysis techniques. A key factor in this equation is that graphs have the unique capability to communicate often complex statistical concepts across numerous speciality areas. No longer are they asked to accept the "mumbo-jumbo" of a poorly understood statistical *t*-test or *p*-value on faith – but, rather, they can now actually see the significance of their data. Simply stated, graphs puts the power of understanding data back into the hands of the individuals responsible for it – the Subject Matter Specialists.

The final step in this implementation strategy is to provide hands-on, 1-on-1 guidance at the user's desktop. (We have found that users often reach dead-ends in the implementation process because of unforeseen hardware, software, and data manipulation problems – and that, although 1-on-1 help is very time consuming, this service is essential to the successful implementation of these new techniques.) The author is also working closely with other individuals in the Bureau charged with key graphical data analysis implementation efforts. For instance, Dave Lassman of Services Division is currently developing new "Front End" software – which presents Analysts with a point-and-click interface – allowing them to easily "slice and dice" large SAS data sets. (Very large data sets are typical of Census data.)

3. NEW EDA GRAPHIC TECHNIQUES

The primary focus of this paper is on a new, very powerful EDA technique – the Industry Profile (IP) Plot. This plot was devised by the author to further promote graphical data analysis at the Bureau. This plot, sometimes quite simple, sometimes combining many variables, is very helpful in that it goes beyond the simple X/Y correlation of two variables. Rather, the fundamental nature of the IP Plot is that it is designed to capture a multivariate correlation of the key variables found on each individual survey form. The IP Plot thus attempts to present a comprehensive overview of each survey variable in a single graph. Often, it is uniquely designed for each type of survey.

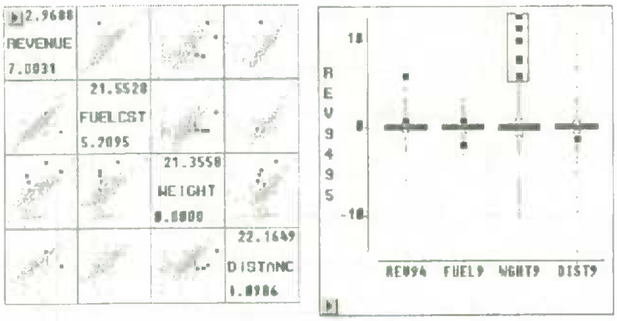
4. KEY TOOLS IN IP MODELING – DATA PROFILE AND LEVERAGE PLOTS

Because an IP Plot is so powerful, we need to exercise caution in its creation. Needless to say, by combining variables to use as ratios in an IP Plot, we first need a very good understanding of the relationships between each of the variables. These relationships may vary markedly for different point cloud clusters (for instance, for industries with different SIC codes), for different geographic areas, or simply for smaller/larger companies. Additionally, time series variances such as business cycles and periodic anomalies could systematically affect these ratios. Basic historical relationships for data sets can vary over time as well. However, by using the power of EDA techniques to help create the IP Plot, these problems can be quickly found and dealt with – and the advantages of this methodology far outweigh these potential problems.

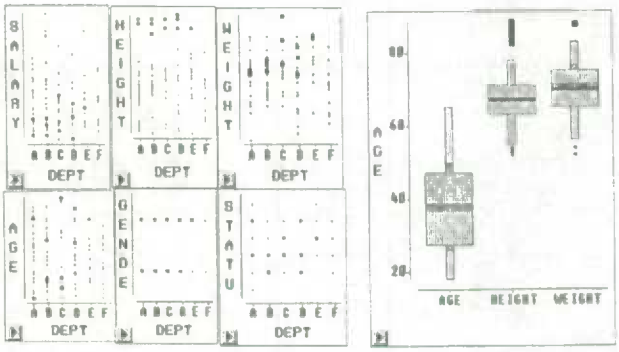
Often sophisticated EDA techniques were used to gain an understanding of the correlation/relationships of all of the key variables of a company/survey. Two of the most useful graphs to us in this task were Leverage Plots and Data Profile (DP) Plots. These two plots were most helpful to us in identifying the covariate influence of each variable, each data cluster, or each individual company in the data set. For further information on Leverage Plots, readers are invited to read the referenced paper by John Sall of SAS Corporation.

The DP plot links a scatterplot matrix with comparative boxplots. Shown below is a DP plot with fictitious data from one of our Transportation Surveys. A key feature of this plot is that all of the key variables can be linked by brushing across each separate graphic window. Thus, if the highest outliers of the box plot for shipping weight are brushed (as is shown here by the darker squares), the corresponding points in the scatterplots are likewise highlighted. Here, for clarity, the scatterplot matrix only shows currently reported Revenues, Fuel consumption, Weights of shipments, and the Distance of goods shipped. (We usually show all of the values reported on the survey form – the reported Maintenance costs, Leasing costs, *etc*.. When required, for closer inspection, users can easily zoom in on any graph.) Since prior year's data is often used as an editing criteria, the box plots show these (current) variables subtracted from their reported values on last year's survey. Thus the outliers of these boxplots quickly show companies with marked year-to-year differences in reported values.

I created this plot to give our Analysts a comprehensive display/summary of all of the variables of each survey form. Given the wealth of information available with this graph (and the power of analysis in this format), the DP Plot has proven itself to be a key tool for editing, exploring data, and gaining unique insights into the interrelationships between survey variables. (I have also formatted a standard set of editing/analysis procedures to use with this graph to gain maximum utility with its use. For instance, both animation and the assignment of progressive color codes to an individual variable have proven to be very effective techniques.) Using this graph, we have also found many previously overlooked problems with our data/methodology – often averaging 1 significant discovery every 15 minutes!



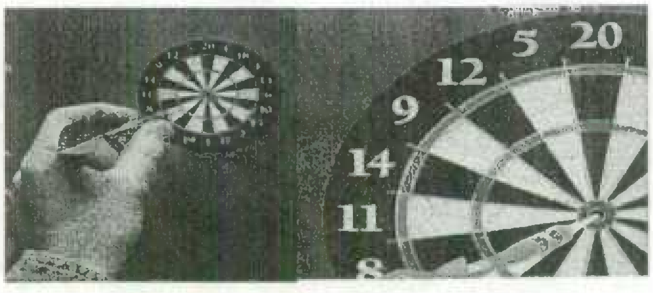
Another example of the power of this graph is shown below. This is a display of the sample data set DEMOG1 that SAS Corporation furnishes with its software. DEMOG1 has been used for years by SAS to teach its traditional Statistical data analysis courses – probably, by now, to thousands of students. Evidently, a great deal of effort has been spent in purposely randomizing this data set. I used a modified DP plot to look at this data set during a break in one of their classes. Because of the power of the DP Plot, I was able to use data mining to quickly spot what would appear to be an apparent case of discrimination against tall employees by some of the departments of this fictitious company! As can be seen below, when we highlight (darker squares) the tallest/outlier employees in the box plot, many of them appear in the lowest rungs of salary for departments A, B, C, & D. We can also see that these employees are randomly distributed by all the other possible explanatory variables (for instance, by age, marriage status, male/female, *etc*.), thus making an interesting basis for a discrimination complaint.

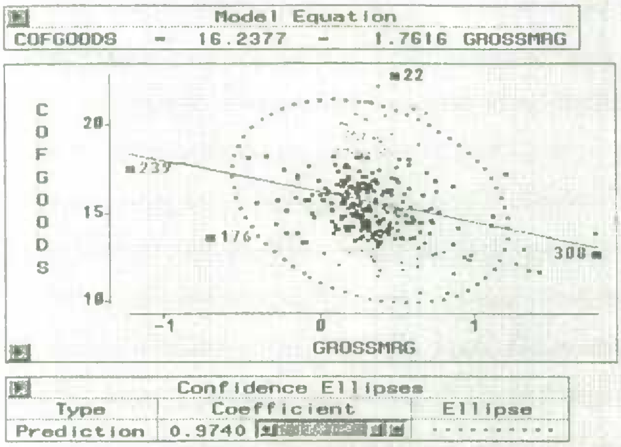


5. EXAMPLES OF THE INDUSTRY PROFILE PLOT

The DP Plot is quite powerful – and, because of this, it can be quite cumbersome/tedious to work with. In contrast, I tried to make the IP Plot as simple to understand/work with as possible.

Some of our Analysts have a hard time understanding some of the fine points of correlations, cluster analysis, and new techniques like leverage plots. However, I've found that by focusing on simplicity (and using the unique communicative/analytical power of graphs as opposed to complex/poorly understood statistical procedures), a number of useful techniques can be implemented. Accordingly, I have found that it often helps to initially explain the concept of an IP Plot in terms of a "shotgun pattern" in a "bulls/eye" type target (shown below). Using this concept, "normal/typical" companies fall near the center of the target/plot – with outlier companies arrayed progressively further from the central point. (Points further and further outside of the center circle/point showing progressively poorer degrees of confidence – less and less normal.) Thus, in the actual IP Plot shown below, the Analyst's attention is drawn to companies # 239, 308, and 22 – they are furthest from the center of the target – and lie furthest outside the 97% confidence ellipse (of a fit between the X & Y axis linked variables).





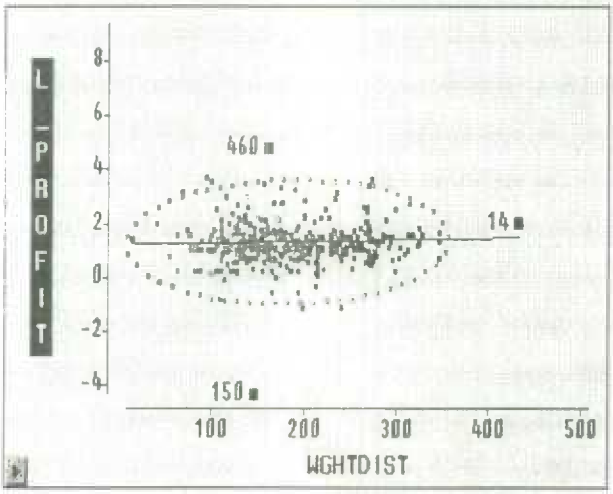
Model Equation

COFGOODS = 16.2377 - 1.7616 GROSSMRG

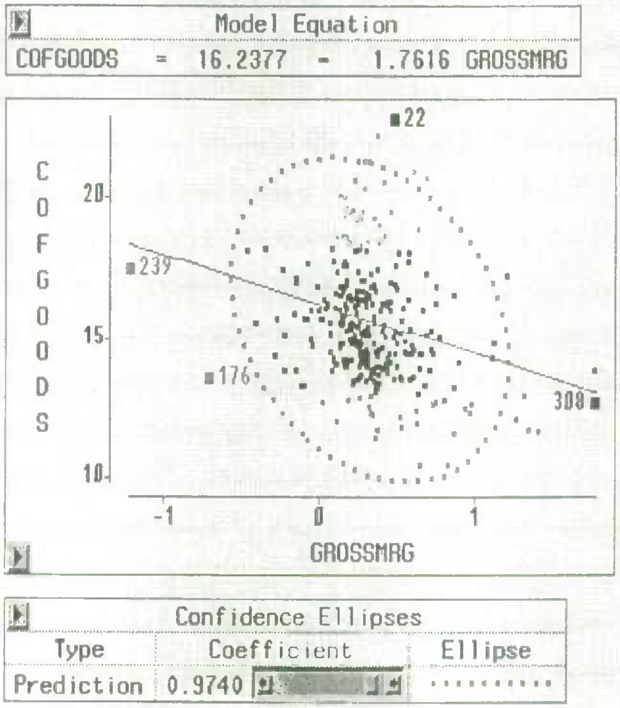
Confidence Ellipses

Type	Coefficient	Ellipse
Prediction	0.9740	....

To further simplify things, concepts like "Profit" *vs* "Work" were derived. For our Transportation Survey, "Profit" would be the reported Revenues, minus all the reported costs – Payroll, Fuel, Maintenance, *etc*. We conceived of "Work" as simply the Weight of a shipment multiplied by the Distance goods were shipped. The resulting plot (with three points that had escaped our analysis to this point) is shown below. (Please note that since the IP Plot displays a large number of interrelated variables from the survey form, it can quickly do of lot of the "pick and shovel" work of an DP Plot. In addition, the correlation line now gives us a level of profit that we look for as "normal" for this type of industry.) Before accepting this plot, I've found that Analysts need to gain a better understanding of the meaning of the points that fall outside of the 95% confidence ellipse. So, I usually introduce the IP Plot in conjunction with the DP Plot and a few simpler plots. Again, using techniques like brushing, they can quickly see/understand/accept what variable(s) contributed to points # 14, #460, and #150 being this far outside the 95% confidence ellipse. Once this basic understanding is gained, it becomes accepted as a standard production/ analysis tool.



Shown below is another example of an IP Plot. Again, a goal of a good IP Plot is to not simply identify key X/Y fits – but to portray a meaningful representation of the key variables. This allows our Analysts to not only edit the real outliers – but to also gain a fundamental understanding of the underlying nature of these data/companies. For our Annual Trade Survey, a "Cost of Goods" and "Gross Margin" terminology was already commonly used. Cost of Goods is simply the current year's Purchases plus draw-down in Inventory over the past year for each company. Gross Margin is reported Sales minus Cost of Goods. These two variables can be thought of as "Expenses" *vs* "Profits". After looking at a number of related industry types with this plot, we were able to verify that these data had a typical elliptical pattern like the profit/work IP plot shown above. After editing a number of outlier points, we found the correlation line in this plot had a clear meaning – a predictable line – which further assisted us in isolating problem companies. (The correlation line connects companies with high expenses and low profits with companies with high profits and low expenses – this makes sense.)



Model Equation

COFGOODS = 16.2377 - 1.7616 GROSSMRG

Confidence Ellipses

Type	Coefficient	Ellipse
Prediction	0.9740	....

6. SUMMARY; PLANS FOR THE FUTURE

Our new DP and IP plots have proven to be very helpful supplements to traditional EDA techniques for the analysis of Bureau data. Likewise, given the success of our initial implementation efforts, our plans for promoting EDA at the Bureau in the future are quite straightforward. We will continue to promote EDA with periodic Bureau seminars covering significant EDA findings. Our current "cookbook" will be supplemented/updated for each new area of the Bureau that implements EDA techniques. We also will continue to repeat the 40-hour introductory EDA course as many times as necessary. (At present, it looks like we will schedule the course exclusively for groups of individuals from a single branch. This will allow students to support one another as they introduce these techniques into their Division.) Again, wherever possible we will offer 1-on-1 hands-on tutoring to individuals in the Bureau. This will allow us to plant EDA "seeds" – and will insure that EDA is used in at least one spot in that Division. (Please note that this work also will often involve customizing our EDA applications to the specialized data within that area – for instance, devising additional unique Industry Profile Plots.)

An on-line Graphics Users Group has also been started. (Every individual in the Bureau is now connected at his/her desktop via networked PCS.) The purpose of this on-line group is to share their data analysis activities, to cover special EDA topics, and to help novice users get thru the rather steep learning curve (and poor documentation) of the SAS software. We will, of course, monitor the progress of these efforts – adjusting where needed. For instance, some of our Analysts are in an often intense Monthly Survey production environment – they have little free time. Given the time requirement of the 40-hour EDA course and the limited time available to them, we are now considering offering a series of 1 day, special topic courses as well. Finally, as a parallel effort, we intend to also work closely with key individuals like Dave Lassman to develop customized user-friendly point-and-click Front Ends for each area/Division of the Bureau.

For further background reading on our efforts to implement EDA at the Census Bureau, readers are invited to read the related paper by Howard Hogan referenced below.

REFERENCES

Bienias, J., Lassman, D., Scheleur, S., and Hogan, H. (1994). Improving outlier detection in two established surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Granquist, L. (1997). Macro-Editing – The aggregate method, Statistical Data Editing, *UN Conference of European Statisticians Statistical Standards and Studies*, Geneva (Switzerland).

Hidirogou, M.A., and Berthelot, J-M. (1986). Statistical editing and imputation for periodic business survey. *Survey Methodology*, 12, 73-83.

Hogan, H. (1995). How exploratory data analysis is improving the way we collect business statistics. *Proceedings of the American Statistical Association*, August 1995.

Fellegi, I.P., and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Sall, J. (1995). Leverage Plots, MS #89115, SAS Institute, Cary, NC.

Tukey, J.W. (1970). *Exploratory Data Analysis*, Addison-Wesley 1, Reading, MA.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

INTEGRATION: THE CENSUS DISSEMINATION PERSPECTIVE

R. Rideout and J. Burgess[1]

ABSTRACT

The challenge of integrating data and products is not a new one for Statistics Canada. Our hundreds of surveys, collecting data across nearly every public and private sector, generates a wealth of information. The possibilities for combining this data across products is perhaps infinite. The challenge has been to develop strategies, methods, and tools that make integration of this data practical to achieve and of value once done.

The Census in many ways is a microcosm of the surveys at Statistics Canada. From a micro database of over 300 variables and over 50,000 standard geographic areas, the Census publishes hundreds of electronic and paper summary data products each Census. Additionally, the Census provides a cost recovery service for custom products, allowing users to define customized products that meet their specific data needs, be they customized themes, variables, or geographic areas. The need to integrate Census data across products is very real for the Census of Population, and the results of that integration can be viewed through the Census product line.

What may not be obvious from a quick view of the product line is how that integration is achieved. Integration of products in the Census begins not with the products, but with the census databases and the tools used to generate products from those databases.

Each Census cycle provides the Census Dissemination Project with the challenge and the opportunity to better integrate it's products. By necessity the Census must continually search for improvements and efficiencies in it's products and the methods used to produce those products. Integration is one Census strategy for achieving it's objectives.

This paper will discuss one of the Census Dissemination solutions to producing integrated products for the 1996 Census.

KEY WORDS: Data integration; Active data dictionary; Metadata.

1. INTRODUCTION

"Canada does not yet have consistent and integrated data to track problems over time, although many of the data needed for new social policy already exist. The difficulty is that the sources cannot be linked since the conceptual framework that would allow their integration has not been formulated."

Some of you may recall hearing Mr. Peter Hicks make this remark during the keynote address at the Statistics Canada Symposium 95. Our colleague Mr. Priest, at the same Symposium, supported this view from the perspective that pressures are growing to build broad based multi source information data bases especially since information technologies are developing which support our ability to do so.

Two years later the debate continues over how statistical agencies can move forward to bring the full range of relevant information into the hands of policy makers, academics, the business community, various levels of government, and the general public.

The Census Dissemination Project Team for the 1996 Census is closely connected with these issues and through this paper will outline a new perspective for exploring the data integration challenge.

2. SURVEY INTEGRATION CHALLENGES

The challenge that Statistics Canada and most other statistical organizations face is to develop links between sources of data and to the extent possible standardize and harmonize any difference in concepts and definitions.

In Figure 1 an example of this definition is presented. In this hypothetical example, 3 different surveys collect information for the same concept – 'family living arrangements'. The definition, for arguments sake, in each case is different. One definition contains lone parents, the others do not; a second contains common-law partnerships, the remaining two do not. In order to integrate the data from all 3 surveys, the more common approach has been to create a single harmonized definition which all 3 surveys agree to adopt at a *particular point in time*. From that point in time the 3 surveys can then move forward using the same 'integrated' concept.

3. CENSUS INTEGRATION CHALLENGES

3.1. The Canadian Census Dissemination Program

Before we look at the Census integration challenges, a quick overview of the Canadian Census Dissemination Program is necessary.

[1] Rick Rideout and Jocelyn Burgess, Dissemination Systems Development Section, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

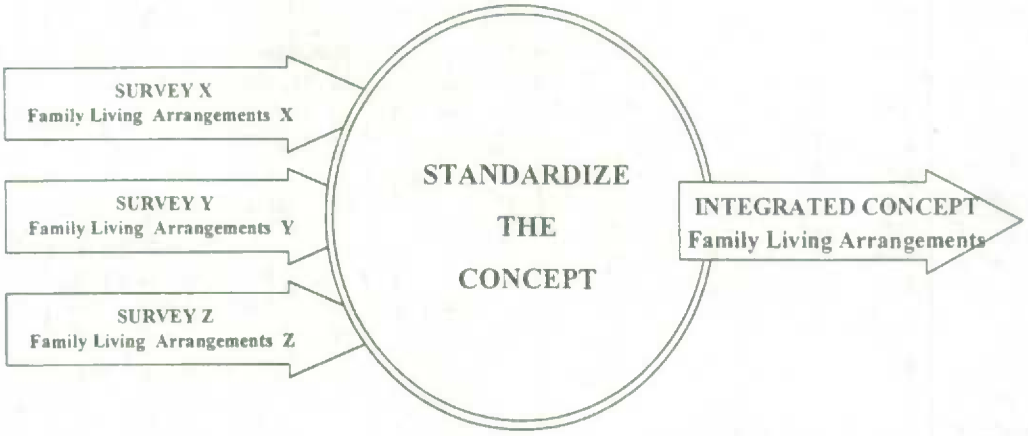

SURVEY X
Family Living Arrangements X
SURVEY Y
Family Living Arrangements Y
SURVEY Z
Family Living Arrangements Z
STANDARDIZE
THE
CONCEPT
INTEGRATED CONCEPT
Family Living Arrangements

Figure 1. Integrating Survey Concepts

The census database is Canada's richest collection of social and economic data. It is also one of the most complex. The census databases contain over 100 gigabytes of microdata distributed over more than 60,000 distinct geographic areas. These data are further classified by over 300 thematic characteristics or variables across 5 major entities: Persons, Households, Dwellings, Census Families, and Economic Families. Add the temporal dimension and we have databases at 5 year intervals from 1971 through 1996. Canadian census databases are further partitioned into 100% (full census) coverage and sample coverage elements (usually a one-in-five household sample).

The 1996 Census product line includes the dissemination of over 2,000 tables on the Internet associated with the release of variables; 24 CD ROM products which contain national and regional profiles, many on a thematic basis; 60 publications of our most popular data and thousands of mass customized diskettes from our electronic warehouse containing information associated with clients' individual needs. A major component of this standard product line contains historical census information. Additionally, the Census provides a cost recovery service for custom products, allowing users to define customized products that meet their specific data needs, be they unique themes, variables, or geographic areas. The need to integrate Census data across products is very real for the Census of Population, since it is one of the few sources that gives a global picture at the local level over time.

3.2 Census Integration Challenges

One of the major challenges for Census dissemination is that Census geographies and the concepts associated with the variables change over time. Boundaries of geographic areas are redefined between censuses as a result of population growth and migration, and the creation of new legal and administrative areas. And as the social and economic fabric of the country changes, so do the variables and the underlying concepts collected, resulting in new concepts or changes in meaning to existing concepts. Some of these differences are obvious to the census user, such as the addition of new concepts or variables. Other differences are more subtle such as the change in meaning of a particular concept.

Nevertheless the Census must be able to provide comparable data across censuses in which the content of the respective databases is often very different. Standardization and harmonization are not options for the Census as the Census must adapt to changing data needs through changes to concepts and at the same time provide historically comparable data. Integration in this context means 'blending' the census concepts across these databases without changing the meaning of the concept in any of the census years. Figure 2. below illustrates this using the Marital Status census concept as an example.

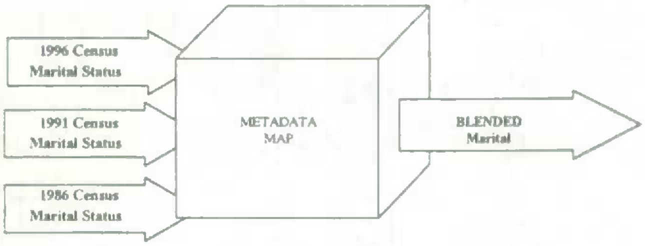


Figure 2. Blending Census Concepts

The Marital Status variable is one example of a concept that has evolved over time. Prior to 1971 the 'Separated' category was not recognized, but rather included in the 'Married' category. And in the 1991 census a second Marital Status variable was added to distinguish Legal Marital Status from Historical Marital Status and in the process recognize 'Common Law' as a marital status category.

In the past, Census relied heavily on highly specialized data analysts who's job (in part) was to know and understand Census concepts, track the changes between Census cycles and produce integrated meaningful products which linked the data over time. This was a daunting task when you consider we needed to track over 60,000 stubs or categories across the more than 300 variables.

Today we have a new approach involving a metadata mapping tool built around a new technology infrastructure which makes this process much easier. However it has taken a lot of hard work to get here. In the early 1990's Census dissemination began to change the way in which we did business. We needed to improve timeliness of delivery of Census products, involve our Regional Offices more directly in the dissemination process, replace aging production tools such as STATPAK and take advantage of new technology. Over the past 7 years we have put in place the foundation which has allowed us to build new automated processes for integrating Census data.

4. THE FOUNDATION FOR CENSUS INTEGRATION

One method we use to accomplish integration is through the use of what we refer to as 'Metadata Maps'. The core of the census dissemination systems is the metadata and it is here that data integration is accomplished. Census metadata is the link between product specifications and the corresponding microdata. Metadata is linked across census years using a 'Metadata Map', essentially a map connecting the concepts across the census years.

Before we look at how the maps are created, an overview of the census dissemination systems architecture is required. It is within this framework that mapping of metadata is accomplished.

The census dissemination systems can be visualized as a 3 tier architecture;

- a microdata layer of census microdata for each quinquennial census from 1971 through 1996.
- a user interface layer consisting of the various application tools used to specify and produce census products.
- a metadata layer containing all of the descriptive metadata required to define microdata, drive application tools, and describe census data and products.

Key to this architecture is the metadata layer consisting of databases of metadata which control all of the specification and production processes associated with producing a census product. There are 3 components on these layers necessary to mapping metadata; an Active Data Dictionary, Electronic Specifications, and Specification Tools.

The 'Active Data Dictionary' on the metadata layer is the data dictionary for the micro databases and is the link to that microdata. The dictionary is considered 'ACTIVE' in that is not simply a repository of metadata stored according to a fixed data model. Rather it is 'active' in that both application tools and users may read and update the content of the dictionary. With an active data dictionary, users can create and store specifications for complex products and can share the specifications or components of specifications across products. And it is flexible, in that new data or business process requirements can be implemented within the dictionary without changing the database model. The Specification Tools are those used by the user to build Electronic Specifications of data products through interaction with the Active Data Dictionary.

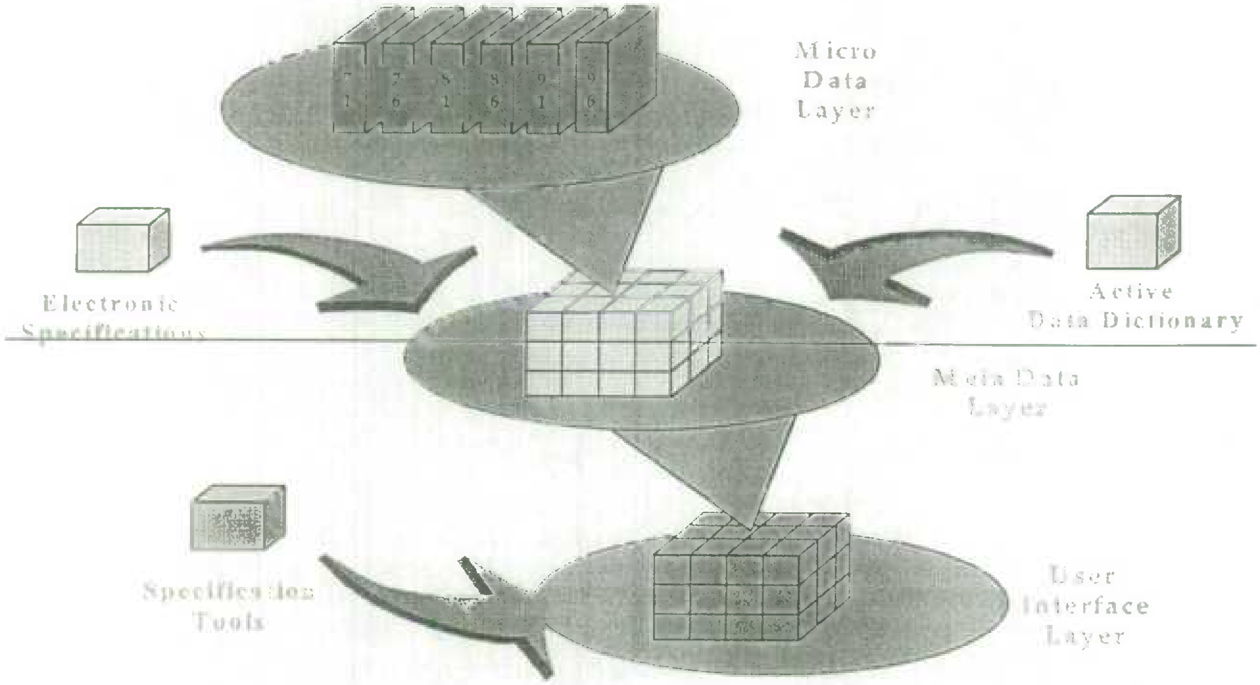


Figure 3. Census Dissemination Systems Architecture

5. MAPPING METADATA

In any census cycle we will specify and produce 1000's of data products. The specifications define the content and structure of the products. Any one product can contain any number of the more than 300 census variables (or characteristics). And there are over 60,000 STUBS (or detailed classifications) across those variables. The solution for the census was to find a method by which we could associate these detailed classifications across time (or census years), in effect build a map of metadata linking variable stubs between census years.

5.1 The Method

Mapping of metadata had to occur at the most detailed level possible to achieve maximum benefit. For us that is the detailed stub level and so we established a 'BRIDGE' between the detailed stubs across census years. On each bridge we placed a 'Metadata Map' establishing the relationship between variables and stubs across the census years. The maps are stored as part of the metadata layer on the databases. In effect, the map is a map between the active data dictionaries of the various census databases. These maps support all possible relationships at the stub level including many to many and none to many.

5.2 The Mapping Tool

A mapping tool was developed that is used to build the metadata maps. The tool provides both automated and manual mapping capabilities, where metadata between censuses is compared and linked in meaning.

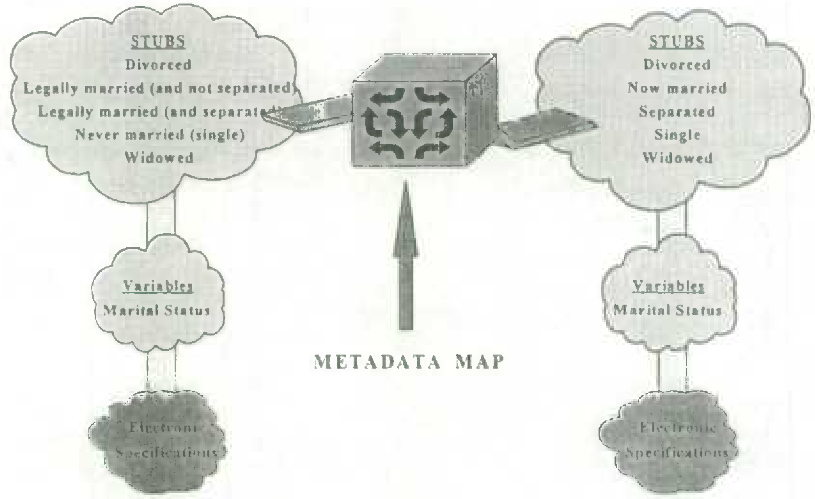
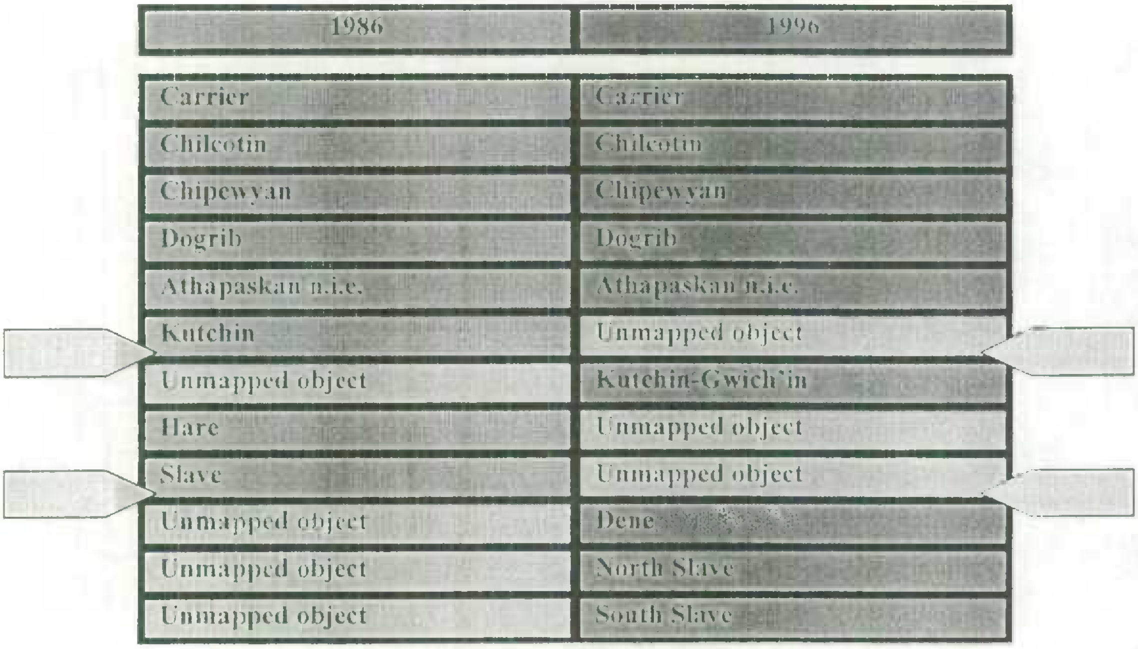


Figure 4. Mapping Metadata



1986	1996
Carrier	Carrier
Chilcotin	Chilcotin
Chipewyan	Chipewyan
Dogrib	Dogrib
Athapaskan n.i.e.	Athapaskan n.i.e.
Kutchin	Unmapped object
Unmapped object	Kutchin-Gwich'in
Hare	Unmapped object
Slave	Unmapped object
Unmapped object	Dene
Unmapped object	North Slave
Unmapped object	South Slave

Figure 5. Automated Metadata Mapping

5.2.1 AutoMap

An AUTOMAP function invokes a comparison of stubs between the 2 databases. Important to note is that the textual labels of the stubs are compared, not the underlying codesets. Codesets do not have to be synchronized across the databases. Relationship between variable values is established using the text labels.

Exact matches are associated in a 1 to 1 relationship. Near matches in stub text are placed near each other in the map, but are not directly associated with each other. Stubs which meet neither the exact or near criteria remain as unmapped objects. Automap is able to establish linkage for most census metadata. This is because most census concepts at the stub level are harmonized across census years.

5.2.2 The Completed Map

Where the stub labels have changed between years manual mapping must be done. To complete and resolve any near match or unmapped objects, stubs in one column can be dragged to the correct location, that is beside the corresponding stub in the other column. More than one stub can be associated with a corresponding stub.

In the example, Chipewyan in 1986 is equivalent to Chipewyan and Dene in 1996. Note also that stubs with different labels can be associated indicating a change in the textual label. Stubs that cannot be mapped because of no associated stub in the other database would remain unmapped. The completed map, once verified, is stored within the metadata layer establishing the linkage of the concept across census years. The maps are bi-directional meaning that they can be applied against data in either direction. For example, the path from 96 concepts to 86 concepts is simply the same path in the opposite direction for 86 to 96. A one to many relationship in one direction becomes a many to one relationship in the other direction.

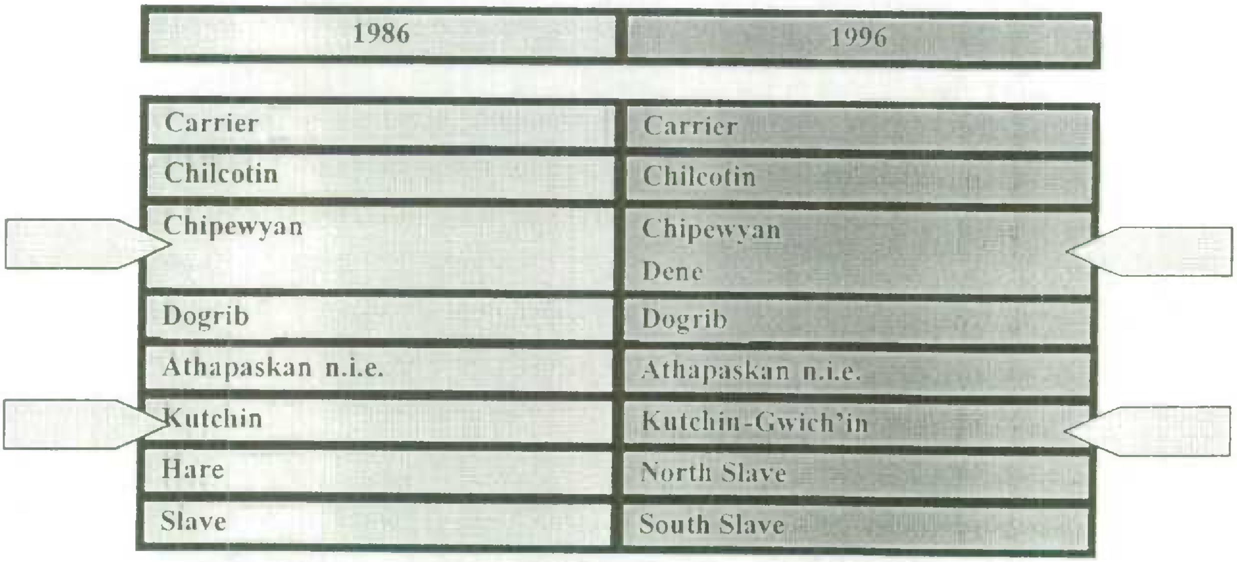
5.2.3 Applying The Map

Once maps are completed product specifications created on one census year can be applied against other census years. This is done using a specification EXPORT tool. The Export tool copies the electronic specification from one census year to another. The copy is performed by sending the specification through the map and adjusting the specification according to the stub association rules in the map.

The copy of the specification is not a physical copy but rather, a copy of the *MEANING* of the specification. For example exporting the 1986 Chipewyan language to 1996 would result in a 1996 specification containing a *Grouping* of Chipewyan and Dene. Data extracted from the 1996 database using the exported specification would be comparable to the data extracted from 1986 using the same specification.

5.3 Some Key Considerations

- This technique of mapping metadata to integrate data would not have been possible without the system architecture census dissemination has in place. An active data dictionary using a metadata driven approach to tools allowed us to build upon the existing systems design and infrastructure. Applications and databases did not have to be redesigned, rather they were extended in functionality and content.
- By removing the need to translate product requirements into codes, users and analysts can focus on the meaning and intent of the census concept. They no longer need be concerned about underlying database codesets, thereby simplifying and reducing the effort required to specify and produce a product.



1986	1996
Carrier	Carrier
Chilcotin	Chilcotin
Chipewyan	Chipewyan Dene
Dogrib	Dogrib
Athapaskan n.i.e.	Athapaskan n.i.e.
Kutchin	Kutchin-Gwich'in
Hare	North Slave
Slave	South Slave

Figure 6. The Completed Metadata Map

- Until we implemented metadata mapping we were dependent upon individual knowledge of census concepts and their structure over time to produce products. We've now captured that knowledge in databases making it readily accessible to anyone working with census data.

6. ONGOING OBJECTIVES

Mapping at the metadata level has as a first objective allowed us to address integration of census concepts across time. And done so while recognizing that the census must continue to provide historically comparable data as well as data reflecting changes to concepts.

As one step towards integration within the census it has also contributed to several other ongoing objectives in census dissemination.

- The Census metadata house is a large one, requiring that the metadata be well structured, organized, and accessible. Metadata maps assist us in structuring, organizing, and linking our metadata.
- Metadata maps are an example of the type of census knowledge that traditionally rested with the individual. Capturing this knowledge through databases centralizes our knowledge base of census and preserves this knowledge for those who will come after us.
- Mapping has allowed us to simplify and improve the processes around providing census products and services by reducing the manual intervention required to produce historically comparable data products.

Through this paper we've presented perhaps a different approach to integration, an approach that allows us to continue to meet client needs for both historically comparable data and data concepts that change over time. We've done this by linking the meanings of census concepts across time.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

THE GIS/MIS DATA MODEL CONCEPT

D.M. Kotun, K.M. Pudlowski[1] and N.K. Arora

ABSTRACT

The concept of combining the strength of the Geographic Attribute Database (GADB) and other existing major databases with a Management Information System (MIS) will greatly improve the ability of management to make timely, effective, and informed decisions. Management will be capable of accessing reports using any one or a combination of databases from one relational database. This paper will describe how using a GIS to incorporate all major databases in the Census will improve data quality and analysis of management reports. Examples of existing MIS reports and proposed GIS/MIS reports on completion status and mailback from a prototype data model will be used to effectively present the GIS/MIS system and its utility as a management reporting tool.

KEY WORDS: Management Information System (MIS); Geographic Information System (GIS); Geographic Attribute Database (GADB); Completion status; Mailback; Relational database.

1. INTRODUCTION

A Geographic Information System (GIS) is a computer-based system that allows the user to focus on the spatial aspects of data in a visual format. Data of various types are then viewed in either a bivariate or multivariate form as maps. The GIS technology has been available in both the public and private sectors for a number of years, however, the development of these systems has now become more cost effective, and GIS technology is more mainstream. It is the authors' goal that this paper will enlighten data collection managers to the capabilities and benefits of utilizing a spatial approach to reviewing data in order to better manage large projects such as the Census of Canada.

2. BACKGROUND AND ASSUMPTIONS

2.1 Assumptions

In order to effectively understand the ideas behind the GIS/MIS Data Model Concept, we can make several assumptions in regards to the 2001 Census of Canada. 1. Geography will remain an integral part of the Census Program. 2. Positions will be staffed based on some level of geography. 3. This workforce will need to be paid (*i.e.*, finance). 4. There will be a logistical component. 5. Direction and control of Census activities will continue to be based on management's use of MIS reporting tools. These assumptions, while generalized, are important in forming the basis for a successful Census. As each of the databases associated with the assumptions has a Geographic element, we propose that a GIS linked with these databases would greatly enhance the performance of the overall database structure.

2.2 Existing Structure

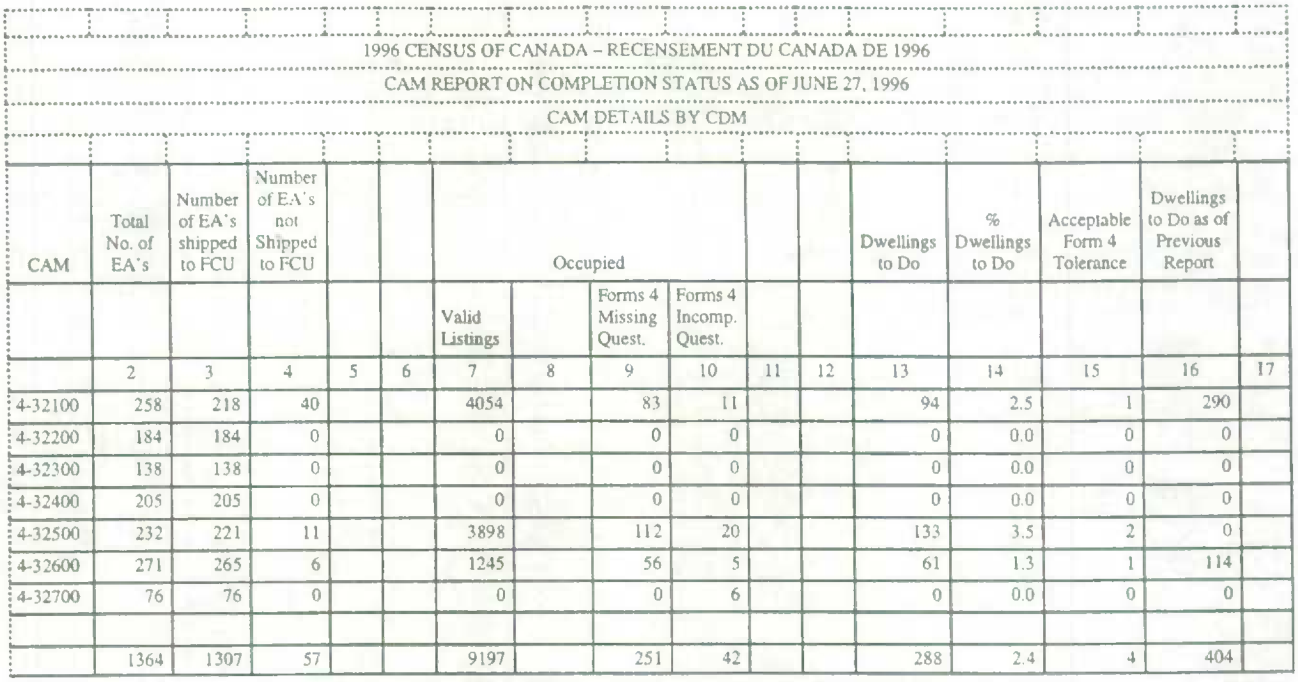
In the 1996 Census of Canada, there were five separate databases utilized for Staffing, Pay, MIS, Logistics, and Geography. While there are some links between tables, vast improvements could be made to this system by combining the information into a series of tables that are interlinked with one another. Under the present system, there are several situations which could potentially cause errors, mismanagement, or impede the smooth completion of the project. Currently, geography updates are forwarded to finance and/or logistics manually.

3. CENSUS COLLECTION DATABASES

3.1 Management Information System Database

A MIS is an information collecting and reporting system designed to give management the information they require to more effectively manage their respective areas of responsibility. During the 1996 Census, MIS provided all levels of management with information required to control and/or modify operations in the field. This information was collected and compiled in the field using a series of reports with differing information. These reports were Census Commissioner (CC) Hiring, Census Representative (CR) Hiring, CR Training, Questionnaire Drop off, Mail Receipts, Completion Status, Quality Control and cost, and pay and expense forms received. Two examples of existing MIS reports are Completion Status by CAM Area and Mail Receipts (see Figure 1 and Figure 2). The accumulated reports are then compiled and sent to the Regional Census Office where they are analyzed by the Census District Manager, Regional Census Manager, and Assistant Director, and Director.

[1] Kent Pudlowski, Informatics, Statistics Canada, Prairie Region, 7th floor, Park Square Bldg., 10001 Bellamy Hill, Edmonton, Alberta, Canada, T5J 3B6.



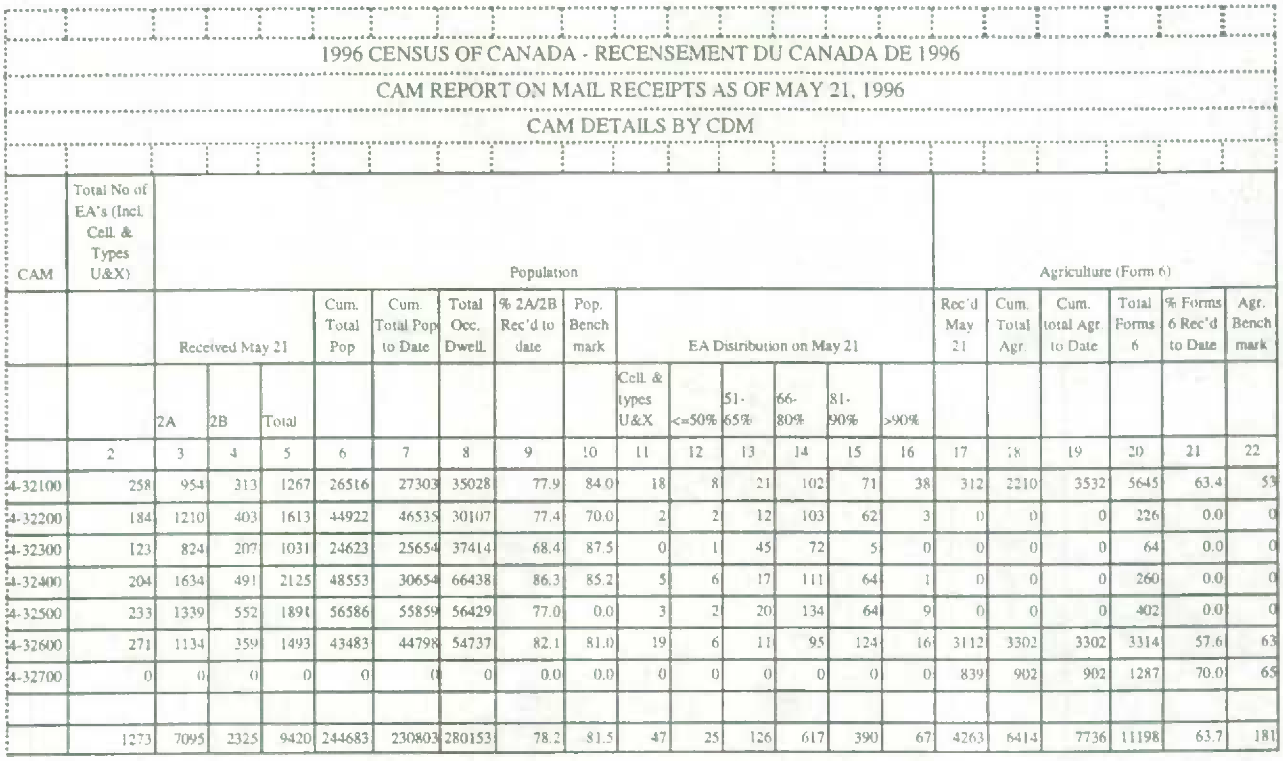
1996 CENSUS OF CANADA – RECENSEMENT DU CANADA DE 1996

CAM REPORT ON COMPLETION STATUS AS OF JUNE 27, 1996

CAM DETAILS BY CDM

CAM	Total No. of EA's	Number of EA's shipped to FCU	Number of EA's not Shipped to FCU			Occupied						Dwellings to Do	% Dwellings to Do	Acceptable Form 4 Tolerance	Dwellings to Do as of Previous Report	
						Valid Listings		Forms 4 Missing Quest.	Forms 4 Incomp. Quest.							
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
4-32100	258	218	40			4054		83	11			94	2.5	1	290	
4-32200	184	184	0			0		0	0			0	0.0	0	0	
4-32300	138	138	0			0		0	0			0	0.0	0	0	
4-32400	205	205	0			0		0	0			0	0.0	0	0	
4-32500	232	221	11			3898		112	20			133	3.5	2	0	
4-32600	271	265	6			1245		56	5			61	1.3	1	114	
4-32700	76	76	0			0		0	6			0	0.0	0	0	
	1364	1307	57			9197		251	42			288	2.4	4	404	

Figure 1. Census MIS Report – Completion Status as of June 27



1996 CENSUS OF CANADA - RECENSEMENT DU CANADA DE 1996

CAM REPORT ON MAIL RECEIPTS AS OF MAY 21, 1996

CAM DETAILS BY CDM

CAM	Total No of EA's (Incl. Cell. & Types U&X)	Population														Agriculture (Form 6)					
		Received May 21			Cum. Total Pop	Cum. Total Pop to Date	Total Occ. Dwell.	% 2A/2B Rec'd to date	Pop. Bench mark	EA Distribution on May 21						Rec'd May 21	Cum. Total Agr.	Cum. total Agr. to Date	Total Forms 6	% Forms 6 Rec'd to Date	Agr. Bench mark
		2A	2B	Total						Cell. & types U&X	<=50%	51-65%	66-80%	81-90%	>90%						
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
4-32100	258	954	313	1267	26516	27303	35028	77.9	84.0	18	8	21	102	71	38	312	2210	3532	5645	63.4	53
4-32200	184	1210	403	1613	44922	46535	30107	77.4	70.0	2	2	12	103	62	3	0	0	0	226	0.0	0
4-32300	123	824	207	1031	24623	25654	37414	68.4	87.5	0	1	45	72	5	0	0	0	0	64	0.0	0
4-32400	204	1634	491	2125	48553	30654	66438	86.3	85.2	5	6	17	111	64	1	0	0	0	260	0.0	0
4-32500	233	1339	552	1891	56586	55859	56429	77.0	0.0	3	2	20	134	64	9	0	0	0	402	0.0	0
4-32600	271	1134	359	1493	43483	44798	54737	82.1	81.0	19	6	11	95	124	16	3112	3302	3302	3314	57.6	63
4-32700	0	0	0	0	0	0	0	0.0	0.0	0	0	0	0	0	0	839	902	902	1287	70.0	65
	1273	7095	2325	9420	244683	230803	280153	78.2	81.5	47	25	126	617	390	67	4263	6414	7736	11198	63.7	181

Figure 2. Census MIS Report - Mail Receipts as of May 21

3.2 Employee Database

The employee database contains all employee information (name, address, phone number, position, *etc.*) along with their assigned geographic location.

3.3 Finance Database

The finance database contains all information regarding appropriate pay to Census Representatives and Census Commissioners in the field. These employees were paid by number of dwellings and size of their assigned land area.

3.4 Logistics Database

The Logistics Database contains all information pertaining to shipping or receiving of supplies and census questionnaires. Bulk shipments of all forms and questionnaires were received and logged into the logistics database. The database then calculated the required amount of supplies for each CAM area based on geographic attribute data from the GADB.

3.5 Geographic Attribute Database (GADB)

Regional offices maintained all geographic attribute data on a computerized system. New enumeration areas could be created and defined according to their specific geographic attributes. Other data stored included land area, dwelling counts, finance and data collection information. This included the province (PROV), federal electoral district (FED), enumeration area (EA), and census collection area, which enabled management to view the data related to any given land area.

4. GEOGRAPHIC INFORMATION SYSTEMS

4.1 Definition

A geographic Information System (GIS) can be defined as:

> *"... a computer based system that provides the following four sets of capabilities to handle georeferenced data: 1. input; 2. data management (data storage and retrieval); 3. manipulation and analysis; and 4. output." [Aronoff, 1991, p.39].*

A GIS could integrate the five aforementioned databases with the common geographic links between them, namely PROV/FED/EA, CC District and CAM area (Figure 3). A series of reports could then be generated from any of the individual databases or a combination of two or more databases that have information at the same geographic level. It is important to note that an update in one database would result in an update in all relevant databases, thus eliminating the need for manual updates in each database.

4.1.1 Input

A basic GIS system is comprised of two data components – a series of graphical features, and the associated attribute data in tabular format. A basic premise of this system is that geographical data must be intelligent, not simply "dumb graphics". Intelligent graphics focus on geographic analysis rather than mere display. Briefly, a drafting or computer mapping system can only produce good graphic output (*i.e.*, maps and pictures), whereas GIS contains a functional database. The concept of a database is essential to a GIS and all contemporary geographic information systems include a database management system. [Antenucci *et al.* 1991, p.87]

For Census purposes, it is not sufficient to have a series of lines that represent roads. These lines must have a form of intelligence including name, type, address ranges, *etc.* Examples of this intelligent Graphical data are, Street Network, Geographical Census Boundaries, Hydrology, and Potential dwelling locations. This type of graphical information is an essential element of the proposed GIS/MIS system as it forms the basic visual component onto which specific data from queried databases are added to.

4.1.2 Data Management (Data Storage and Retrieval)

The attribute data are stored in a series of relational tables as part of a Relational Database Management System (RDBMS). A relational database is a "*...method of structuring data as collections of tables that are logically associated to each other by shared attributes. Any data element can be found in a relation by knowing the name of a table, the attribute (column) name, and the value of the primary key*". [Environmental Systems Research Institute Incorporated, 1995, G18].

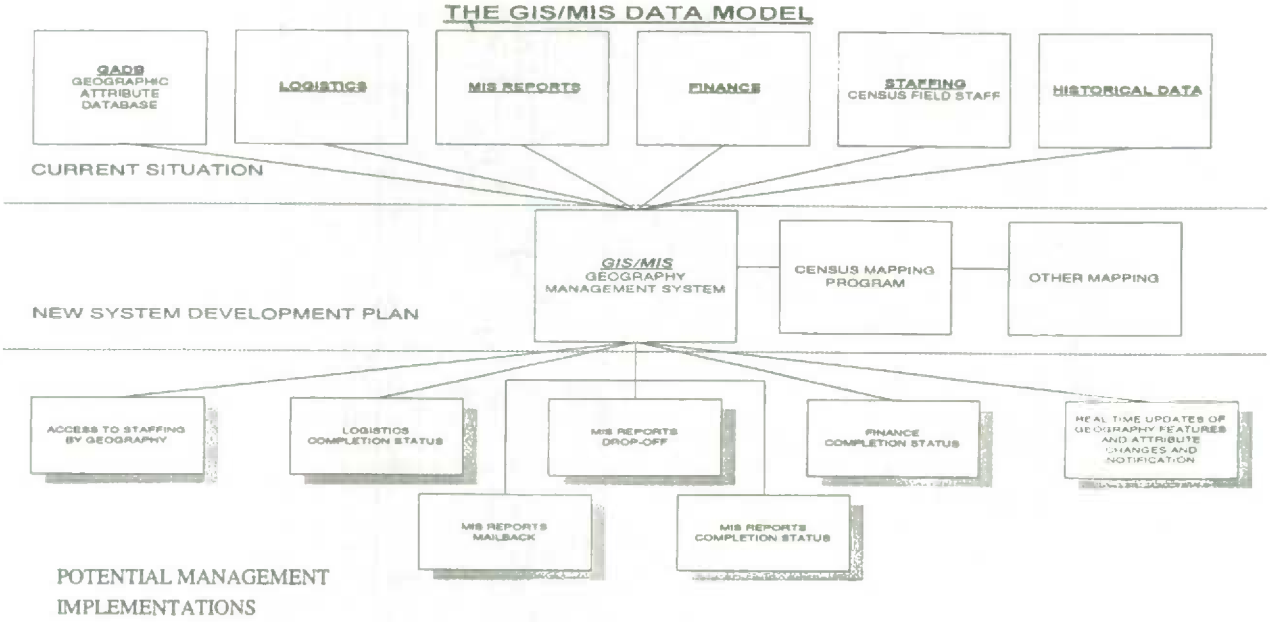

THE GIS/MIS DATA MODEL
GADB
GEOGRAPHIC ATTRIBUTE DATABASE
LOGISTICS
MIS REPORTS
FINANCE
STAFFING
CENSUS FIELD STAFF
HISTORICAL DATA
CURRENT SITUATION
GIS/MIS
GEOGRAPHY MANAGEMENT SYSTEM
CENSUS MAPPING PROGRAM
OTHER MAPPING
NEW SYSTEM DEVELOPMENT PLAN
ACCESS TO STAFFING BY GEOGRAPHY
LOGISTICS COMPLETION STATUS
MIS REPORTS DROP-OFF
FINANCE COMPLETION STATUS
REAL TIME UPDATES OF GEOGRAPHY FEATURES AND ATTRIBUTE CHANGES AND NOTIFICATION
MIS REPORTS MAILBACK
MIS REPORTS COMPLETION STATUS
POTENTIAL MANAGEMENT IMPLEMENTATIONS

Figure 3. The GIS/MIS Data Model

As discussed under description of Census Databases, the proposed system will require a series of relational tables (Figure 4).

As depicted in Figure 4, all 5 databases in the proposed GIS/MIS system are related. The key component or link for the tables is the actual ground location as defined by the PROV/FED/EA number. Each of these PROV/FED/EA numbers are unique and defined by an 8 digit identifier. This identifier consists of a two digit PROV Code, a 3 digit FED Code, and a 3 digit EA Number.

4.1.3 Data Manipulation and Analysis

Presently, management is spending an excessive amount of time trying to relate this information by analyzing these series of tables without a visual component. This is a very inefficient and ineffective use of management resources. For example, there are roughly 10,000 EA's which define the Prairie Region. It is virtually impossible for a manager to study over 10,000 lines of data in a timely fashion and arrive at a basic understanding of the results, let alone implement any preventative and/or corrective measures. A visual component would enhance the ability of management to make timely, informed, accurate, and effective decisions. This is the foundation of our proposed system.

There are five basic categories of queries management can make to answer questions relating to the Census. These are:

(i) Location – These are specific queries regarding a particular geographical location. *Is this a bilingual canvasser EA?*

(ii) Condition – Queries under this section are not confined to specific in location, but locate areas where certain conditions are satisfied. *Where are all EA's with completion status below the benchmark?*

(iii) Trends – Queries of this type are temporal in nature. Compare total cost per unit between Census periods. *Have preventative actions increased the completion status of targeted areas?*

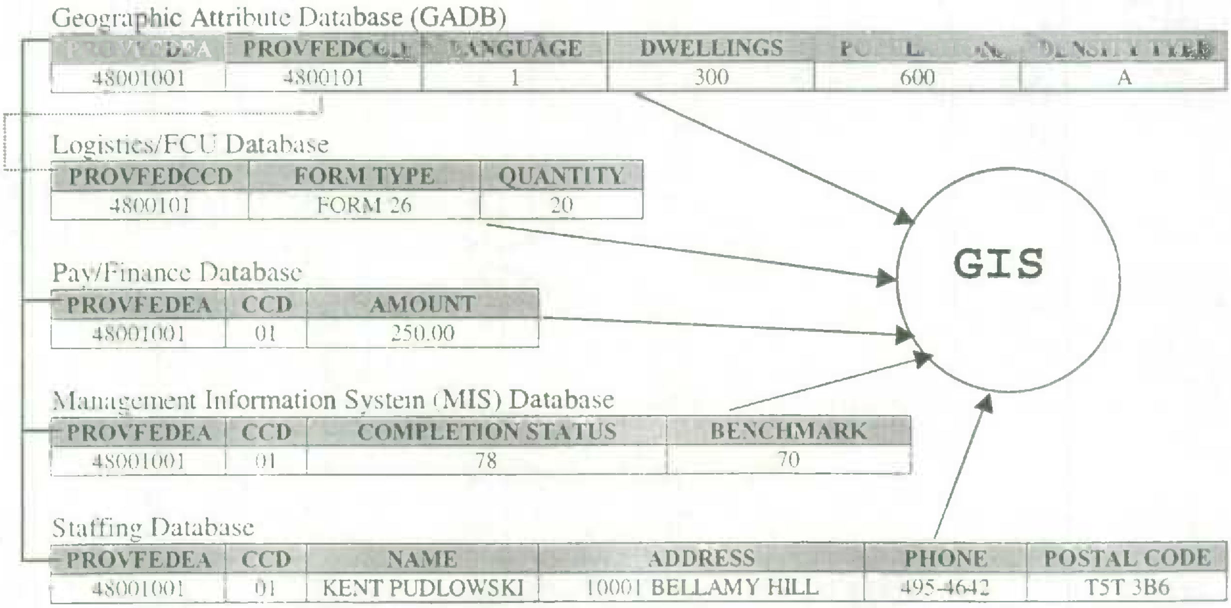
(iv) Patterns – These are extremely focused and complex queries that may combine information from a series of databases. Under the current system, it is impossible to accurately assess these types of queries, and even more impossible to capitalize on a positive trend or eliminate a negative one. *Show canvasser residential areas that have submitted mileage claims above a certain amount and have a completion of less than 98%?*

(v) Modeling – These queries measure the effects of altering an operational situation. *Where are the optimum training locations that minimize travel expenses? What are the optimum training locations that minimize travel expenses?*

5. PROTOTYPE GIS/MIS DATA MODEL

During the 1996 Census of Canada, employees of the Prairie Region created a rudimentary prototype GIS/MIS system. This system focused on creating a series of basic maps on mailback and completion status.

There are a series of inherent inaccuracies in the prototype and implied methodology due to data limitations. However, regardless of these limitations, this model shows this system is possible and the methodology correct with complete data. Overall, management found that the utilization of these thematic maps to minimize the time and effort spent reviewing countless GIS and MIS reports was effective.



Geographic Attribute Database (GADB)

PROVFEDEA	PROVFEDCCD	LANGUAGE	DWELLINGS	POPULATION	DENSITY TYPE
48001001	4800101	1	300	600	A

Logistics/FCU Database

PROVFEDCCD	FORM TYPE	QUANTITY
4800101	FORM 26	20

Pay/Finance Database

PROVFEDEA	CCD	AMOUNT
48001001	01	250.00

Management Information System (MIS) Database

PROVFEDEA	CCD	COMPLETION STATUS	BENCHMARK
48001001	01	78	70

Staffing Database

PROVFEDEA	CCD	NAME	ADDRESS	PHONE	POSTAL CODE
48001001	01	KENT PUDLOWSKI	10001 BELLAMY HILL	495-4642	T5T 3B6

Figure 4. Detailed View of the Proposed GIS/MIS Data Model

The Completion Status by CAM Area (Figure 5) shows the remaining completion percentage. The dark polygons show areas near completion, therefore, the area is above the local benchmark. The light areas show locations in need of improvement. Almost immediately management can make decisions based on this figure and take appropriate action

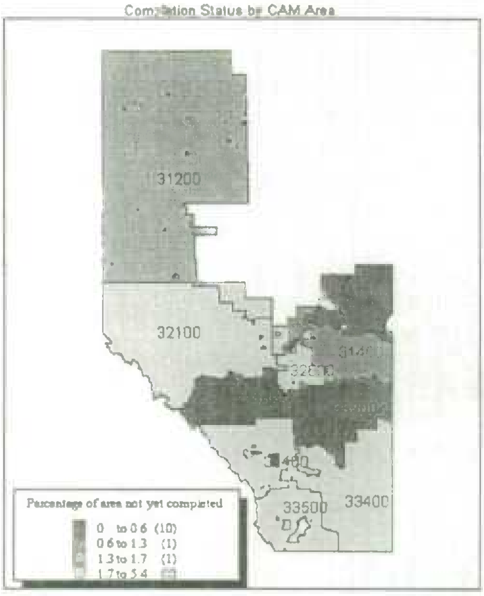


Figure 5. Census GIS Report – Completion Status by CAM Area

As shown by Figure 6, Mail Receipts differed by areas within the province. With the use of this "visual report", management can quickly assess areas with lower mail receipts and implement contingency plans. Therefore, management can be proactive and may be able to make corrective decisions based on these new reporting tools. (NOTE: Complete data was not used for the examples portrayed in Figures 5 and 6 in order to maintain simplicity in the model. Therefore, certain geographical regions are not displayed in the examples.)

6. CONCLUSION

The information presented in this paper provides better insight into how Geographic Information System and Management Information Systems can work together. When these two systems are combined, the result is a data based visual array of analysis and reporting information. This GIS/MIS Data Model can combine all previous Census databases and is based on relational links between them. This will ensure reliability and accuracy of all data that passes through it, thus allowing management to capitalize on a positive trend or act to minimize a negative one. Management can then make timely and informed decisions in an effective and efficient manner. The functions and abilities of the GIS/MIS data model concept would be invaluable to the Census of Canada, as well as to other related geographic based projects.

Mailback Mail Receipts by CAM Area

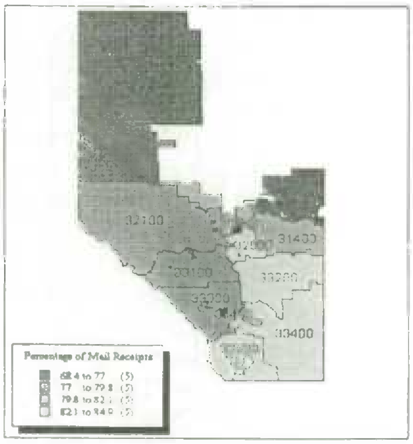


Figure 6. Census GIS Report – Mail Receipts by CAM area

REFERENCES

Antenucci, J.C., Brown, K., Croswell, P.L., Kevany, M.J., and Archer, H. (1991). *Geographic Information Systems – A guide to the Technology*. Chapman and Hall, New York, New York.

Aronoff, S. (1991). *Geographic Information Systems: A Management Perspective*. WDL Publications, Ottawa, Ontario.

Environmental Systems Research Institute Inc. (1995). *Understanding GIS – The ArcInfo Method*. Geo Informational International. Third Edition, New York: John Wiley & Sons.

SESSION C-6

Sample Design and Sample Selection Methodologies

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

DETERMINING THE PROBABILITIES OF SELECTION IN A MULTIVARIATE PROBABILITY PROPORTIONAL TO SIZE SAMPLE DESIGN

J.F. Amrhein, C.M. Fleming and J.T. Bailey[1]

ABSTRACT

Hicks *et al.* (1996) presented a technique for drawing a PPS sample across multiple list frames for a multi-purpose survey that had desirable properties under a reasonable set of models (one for each purpose). In their technique the frames are not stratified and each frame only consists of units which possess a specific item of interest. A unit may exist in more than one frame depending on the number of items it possesses. The technique determines a unit's probability of selection by taking the product of the number of draws, m, and the maximum over all frames of the unit's univariate relative measure of size. A part of the technique involves finding the *m* that meets item-level target sample sizes while minimizing the total sample size. In another design, Bailey and Kott (1997) describe a technique where *m* is unique for each frame and a unit's final probability of selection is the maximum over all frames of the unit's frame-specific product of *m* and the relative measure of size. This paper describes a technique for finding the optimum probability of selection under the same models and constraints motivating Hicks *et al.* and Bailey and Kott. In this alternative method, the probability of selection is found by solving a convex programming problem. The numerical solution is obtained by using Chromy's algorithm. The new method is applied to the data sets for which the original designs were developed and the results compared.

KEY WORDS: Effective sample size; Optimal selection probabilities; Multipurpose survey; Convex programming.

1. INTRODUCTION

The National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA) is experimenting with new sample designs for two of its multi-purpose surveys. The Vegetable Chemical Use Survey (VCUS) uses a two-phase design in which data for acres planted to targeted vegetables is collected in the first phase and acres receiving chemical treatment and amounts applied are collected in the second phase. Estimates from the survey include, by vegetable and active ingredient, percent of planted acres receiving treatment, pounds of active ingredient applied per acre per application, total pounds applied per acre and total pounds applied. Details for the new design and its development can be found in Hicks *et al.* (1996). The second survey is the Crops Survey (CS) which measures acreage, yield, production and stocks for many different crops. This survey is conducted quarterly and the targeted crops change by quarter. Additionally, some crops are very important, but rare, and so require special attention. Details for this design can be found in Bailey and Kott (1997) . The probabilities of selection for these surveys rely on a relative measure of size for each population unit which is determined by the acreage of each targeted crop that a unit possesses. A systematic PPS technique is used to select the VCUS sample in the second phase and a Poisson technique is used for the CS.

This paper describes an alternative method of determining the probabilities of selection. In this approach, the probabilities are found by solving a convex programming problem using Chromy's (1987) algorithm. In the next section we discuss our methods for finding the probabilities of selection under the current design and using the proposed technique. We also discuss our methods of comparison. The final section presents results from our comparison and some final remarks.

2. METHODS

We assume the following model at the sample design stage.

$$y_{k,i} = \beta_k x_{k,i} + x_{k,i} \epsilon_{k,i} \qquad (1)$$

where k represents a targeted crop, i represents a population unit and $\epsilon_{k,i}$ is a random error term with mean zero and variance $\sigma^2_{k,i}$. For the VCUS, the dependent variable y is chemical use and the independent variable x is planted acreage reported during the first phase. For the CS, y is the survey response of acreage and x is the acreage maintained on the list sampling frame. We also assume that $\sigma^2_{k,i} = [c_k\, x^{\gamma-1}_{k,i}]^2$ and define

$$p_{k,i} = \frac{f_i x^{\gamma}_{k,i}}{\sum_{i \in F} f_i x^{\gamma}_{k,i}}$$

[1] John F. Amrhein, Charles M. Flemming and Jeffrey T. Bailey, National Agricultural Statistics Services, Room 4818 South Building, Washington, D.C., U.S.A. 20250.

where F is the set of all sampling population units and f_i is the first phase weight for the VCUS and equals 1 for the CS. If the probability of selection of unit i, $\prod_i$, is proportional to $p_{k,i}$, then $\prod_i$ is also proportional to $x_{k,i}\,\sigma_{k,i}$. This set of probabilities of selection achieves asymptotic optimality as defined by Brewer (1963). For the VCUS, $\gamma = 1$ because $r_{k,i} = y_{k,i}/x_{k,i}$ is a present-time ratio, chemical use per planted acre, so it is reasonable to assume that $\mathrm{Var}(r_{k,i}) = \sigma_k^2$ is constant for a given crop. For the CS, $r_{k,i}$ is the ratio of a present value to a frame value and $\gamma = 3/4$ is believed to be more appropriate.

Hicks *et al.* discuss the concept of an effective sample size, n_k^*, which is defined here in a slightly modified form as the ratio of the population variance of $r_{k,i}$ and the anticipated mean squared error of $\hat{r}_k$ under the model. For the case where $\gamma = 1$, effective sample size is:

$$n_k^* = \frac{\sigma_k^2}{E[\hat{r}_k - R_k]^2}$$

where R_k is the true ratio of y to x and $\hat{r}_k$ is a design consistent estimator for this ratio. This quantity can be interpreted as the sample size, under an infinite population, that achieves the anticipated mean squared error of an estimator if asymptotically optimal selection probabilities are used (Kott 1997). Observe that effective sample size is inversely related to anticipated mean squared error: when one doubles, the other is cut in half. Note also that effective sample size is determined for each item of interest in the survey, in our case, for each crop. Therefore, effective sample size provides a measure of how well a design is expected to perform for each item of interest.

It can be shown that for the model in (1):

$$n_k^* \approx \frac{1}{\sum_{i\in F} p_{k,i}^2 \left(\frac{1}{\prod_i} - \frac{1}{f_i}\right)} \tag{2}$$

Note that when all of the $\prod_i$s are small and proportional to $p_{k,i}$ (the value that minimizes the anticipated mean squared error of $\hat{r}_k$) the effective sample size for k is (approximately) equal to the realized sample size. By "realized sample size" we mean the usual definition of sample size, the number of units to be contacted in the survey. This also refers to crop level sample sizes, n_k, the number of units in the sample possessing crop k.

The problem, then, is to find an appropriate set of $\prod_i$s for sampling. For the VCUS, Hicks *et al.* compute the $\prod_i$s such that $\prod_i = \min(1, m \cdot \max_{1\le j\le K}(p_{j,i}))$ where

$$m \ge \frac{n_k \cdot \sum_{i\in F} \frac{p_{k,i}^2}{\max_{1\le j\le K}(p_{j,i})}}{1 + n_k \cdot \sum_{i\in F} \frac{p_{k,i}^2}{f_i}} \qquad \forall\ k = 1, \ldots, K$$

where n_k is the desired realized sample size for crop k in the finite population and K is the number of crops in the survey. For the CS, Bailey and Kott use $\prod_i = \min(1, \max(n_1 \cdot p_{1,i}, \ldots, n_k \cdot p_{K,i}))$. This is likely to produce expected realized sample sizes of $n_1, n_2, \ldots, n_k$ or greater, but may not (problems arise when some $\prod_i$s are truncated at 1). The expected realized total sample size is $E[n] = \sum_{i\in F}\prod_i$. The expected realized sample size for a particular crop is $E[n_k] = \sum_{i\in F} I_{k,i} \cdot \prod_i$ where $I_{k,i} = 1$ if farm i has crop k and 0 otherwise.

We will refer to these methods as the "maximum" methods and will refer to the $\prod_i$s defined under these methods as the maximum $\prod_i$s since they rely on a unit's largest relative measure of size. This is to differentiate them from the proposed method which we will refer to as the "optimal" method and $\prod_i$s under this method as optimal $\prod_i$s since they result from solving an optimization problem.

In this paper we formulate the problem of finding the $\prod_i$s that minimize the expected realized sample size while meeting or exceeding all target effective sample sizes as a convex programming problem. We can state this formally as:

$$\begin{gathered} \min \sum_{i\in F} \prod_i \\ \ni \\ \frac{1}{\sum_{i\in F} p_{k,i}^2 \left(\frac{1}{\prod_i} - \frac{1}{f_i}\right)} \ge n_k^* \\ 0 < \prod_i \le 1 \end{gathered} \tag{3}$$

Each constraint corresponds to a specific crop so that the optimization is performed for all k where k ranges from 1 to K. The left hand side of a sample size constraint represents the effective sample size under the optimal $\prod_i$s whereas the right hand side is a targeted effective sample size. For purposes of this investigation, the right-hand side is the effective sample size corresponding to the set of maximum $\prod_i$s. An interpretation would then be, "minimize the total realized sample size so that the effective sample size is at least as large as the effective sample size under the current method". If the optimal method generates selection probabilities that sum to a lower overall sample size, then we can conclude that the maximum method generates sub-optimal selection probabilities in the sense that a different set of $\prod_i$s can achieve the same precision for a lower cost (or better precision for the same cost).

To the end of finding a numerical answer, it is easier to work with an equivalent formulation of the convex programming problem. Let $\prod_i = 1/z_i$ and, after rearranging terms in the constraints, the problem can be stated as:

$$\begin{gathered} \min \sum_{i\in F} \frac{1}{z_i} \\ \ni \\ \sum_{i\in F} p_{k,i}^2 z_i \le \frac{1}{n_k^*} + \sum_{i\in F} \frac{p_{k,i}^2}{f_i} \\ z_i \ge 1 \end{gathered}$$

The Kuhn-Tucker conditions imply that the optimum z_is must satisfy $\nabla F = -\sum_k \lambda_k \nabla f_k$ and $\lambda_k f_k = 0\ \forall\ k$ where

$F = \sum_{i \in F} 1/z_i$ and $f_k(z) = \sum_{i \in F} p_{k,i}^2 z_i - (1/n_k^* + \sum_{i \in F} p_{k,i}^2 / f_i)$. By doing the differentiation and equating corresponding elements of the gradients, it can be shown that:

$$\frac{1}{z_i} = \prod_i = \sqrt{\sum_{k=1}^{K} \lambda_k p_{k,i}^2} \qquad (4)$$

Suppose we let $\lambda_k = 1$, then we will obtain a value for $\prod_i$. It will not be the optimum $\prod_i$. A better value can be obtained by letting $v_k(\prod) = \sum_{i \in F} p_{k,i}^2 / \prod_i$, the left hand side of a constraint, and $v_k^0 = 1/n_k^* + \sum_{i \in F} p_{k,i}^2 / f_i$, the right hand side of a constraint. We revise the value of λ_k by letting

$$\lambda_k^* = \lambda_k \left(\frac{v_k}{v_k^0} \right)^2 \qquad (5)$$

and compute a new $\prod_i$, call it $\prod_i^*$ using equation (4).

With this new value of $\prod_i$, λ_k once again can be revised by using equation (5). This time v_k^0 stays the same, but v_k will have been given a new value, $v_k = v_k(\prod^*)$. The process of revising λ_k, calculating $\prod_i$, and assigning a new value to v_k continues until $\lambda_k^* \cdot v_k(\prod^*) = 0$. This iterative process is simply the Chromy algorithm (Chromy 1987; Zayatz and Sigman 1995).

3. RESULTS

To compare the current and proposed methods of calculating selection probabilities, we first set desired realized sample sizes for each crop, then, using equation (2), we calculated effective sample sizes and used these to generate optimal selection probabilities using the inequality in (3).

We performed this analysis for the operational sample from the 1996 VCUS. This survey has a small sampling population for most targeted crops and a complete enumeration of all units possessing these crops is usually desired. Therefore, under the maximum method, many units receive a selection probability of 1. We tested the optimal technique as described above in order to compare sample sizes using $E[n] = \sum_{i \in F} \prod_i$. Figures 1 and 2 show the relationship between the maximum and optimal $\prod_i$s for two of the states in the survey (out of a total of 12 states). Figure 1 shows the relationship for New Jersey as an example of a typical result for the states with many certainty units, although most other states showed less dispersion from the $Y = X$ line. Figure 2 shows the results for Minnesota which targeted only two crops whose populations were large enough that a complete enumeration was not desired. Table 1 displays the expected realized sample sizes.

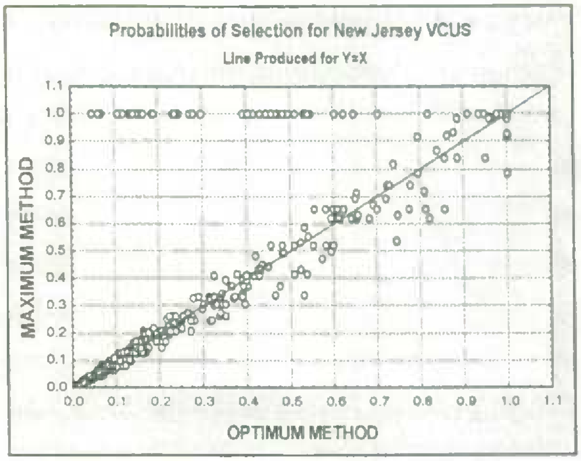


Figure 1. Typical Relationship Between the Probabilities of Selection

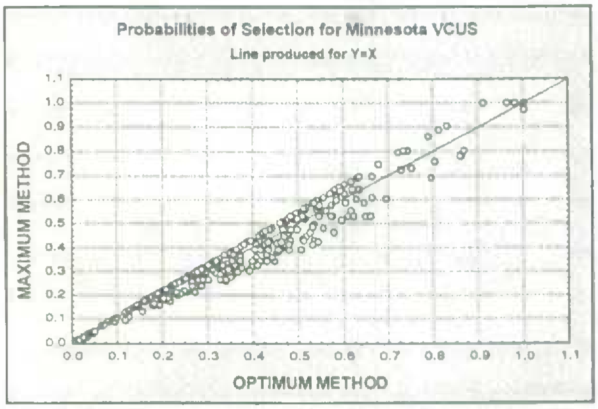


Figure 2. Relationship Between Probabilities of Selection for a State Without Certainties

Table 1
Comparison of Expected Overall Realized Sample Sizes for the VCUS

State	Maximum Method	Optimal Method
AZ	113	103
CA	1,005	840
FL	526	458
GA	495	388
MI	492	440
MN	169	167
NJ	301	272
NY	638	608
NC	689	651
OR	448	425
TX	378	344
WA	404	380
WI	451	413

From these results we can see that some savings in overall sample size can be realized when the aim of the statistician is to target effective sample sizes. The optimal technique does not, in general, select all units with a crop even when the targeted effective sample size was calculated using all available units. The maximum method, targeting realized sample sizes, assigns $\prod_i$s at 1 regardless of a unit's size if a complete enumeration is desired. This approach tends to over-sample small operations. The optimal method does not require all units to reach the targeted effective sample sizes. In essence, the small units do not add to the effectiveness of the sample. However, complete enumerations are desired since estimation is at the crop/active ingredient level and the use of some pesticides is difficult to find.

We also compared the techniques through simulation. For both the CS and the VCUS we expanded survey respondents into populations using the sampling weights to replicate observations. We generated maximum and optimal selection probabilities using the method described above and then selected 1,000 samples using each set of $\prod_i$s. We estimated totals for the CS and a ratio for the VCUS for the 1,000 samples from each method and calculated the root mean squared error (RMSE) and the ratio of optimal RMSE to maximum RMSE.

For the CS two different samples were selected. The first, for row crops, came from a population of about 35,000. The ratio of RMSEs for corn, soybeans, dry beans and sunflowers were 0.894, 1.015, 1.126 and 1.020 respectively. The second sample, for small grains, came from a population of about 30,000. The ratios of RMSEs for spring wheat, winter wheat, barley and oats were 1.002, 1.004, 1.003 and 1.000 respectively. Sample sizes were virtually identical and so we conclude that, for the CS, there appears to be no advantage to generating optimal $\prod_i$s.

For the VCUS, we estimated the application rate per planted acre for the most common active ingredient for each vegetable. Four states were analyzed generating 29 estimates. Results support those from the CS. Ratios of RMSEs ranged from .44 to 3.0 with and average of 1.10 and first, second and third quartile values of .89, 1.03 and 1.11. These results are not surprising given that we did not see a reduction in expected realized sample sizes. If we would have observed a reduction in expected realized sample sizes, then RMSE ratios of 1 would confirm no loss in precision.

Based on our results, we conclude that the maximum methods find near-optimal selection probabilities and do not recommend changing to the optimal method for existing surveys that target realized sample sizes. However, as seen in the results from the operational VCUS comparison, if a survey targets effective sample sizes, which is equivalent to targeting coefficients of variation, then the optimal method will produce a set of selection probabilities that minimizes the overall expected realized sample size while meeting or exceeding the targeted effective sample sizes.

ACKNOWLEDGEMENTS

The authors wish to thank Phil Kott for his reviews of this paper and his helpful comments and advice.

REFERENCES

Bailey, J., and Kott, P. (1997). An application of multiple list frame sampling for multi-purpose surveys. *ASA Proceedings of the Section on Survey Research Methods*.

Brewer, K. (1963). Ratio estimation and finite population: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 5-13.

Chromy, J. (1987). Design optimization with multiple objectives. *ASA Proceedings of the Section on Survey Research Methods*.

Hicks, S., Amrhein, J., and Kott, P. (1996). Methods to meet target sample sizes under a multivariate pps sampling strategy. *ASA Proceedings of the Section on Survey Research Methods*.

Kott, P. (1997). The effective sample size of a ratio or calibrated expansion estimator. Personal communication, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C.

Zayatz, L., and Sigman, R. (1995). CHROMY_GEN: General-purpose program for multivariate allocation of stratified samples using chromy's algorithm. *Economic Statistical Methods Report Series ESM-9502*, Bureau of the Census, U.S. Department of Commerce, Washington, D.C.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

RATE OF CONVERGENCE OF AN ASYMPTOTIC VARIANCE FOR A πps SEQUENTIAL SAMPLING

Y.G. Berger[1]

ABSTRACT

In survey sampling, the variance estimation requires heavy calculation because of the large number of second-order inclusion probabilities. If we implement the Chao sampling (1982), we will see that it is rather easy to compute the variance estimator. The normality of the Horvitz-Thompson estimator is also discussed. In this contributed paper, the results of Berger (1996, 1998a, 1998b, 1997c and 1997d) are presented.

KEY WORDS: Chao sampling; Horvitz-Thompson estimator; Inclusion probabilities; Sampling without replacement.

1. INTRODUCTION

For many users, the need to draw a sample with unequal probabilities might be more common than the need to select a simple random sample. A usual problem in survey sampling is to use a sampling design that allows for unbiased variance estimation. Furthermore, the items should be selected without replacement through a sequential run of a computer file whose length N.

The most common sequential sampling design is the systematic procedure (Madow 1949). Nevertheless, this method has the disadvantage of not ensuring unbiased variance estimation. It also needs a preliminary pass through the file to determine N. The Chao sampling (1982) provides a satisfactory solution. This method is sequential and ensures unbiased variance estimation. Moreover, N does not have to be known before the selection of the sample.

The variance estimator is hard to compute as it needs a cumbersome calculation of a large number of second-order inclusion probabilities. In this paper, we will see that it is possible to use a variance approximation that does not require second-order inclusion probabilities. We will show that the Central Limit Theorem remains valid. Thus usual inference can be made using the normal distribution.

2. NOTATION

Let U be a finite population of N units labelled $i = 1, ..., N$. A sample S is a sub-set of n distinct units from U, where n is the sample size. n is a constant known before the selection of the sample. A sampling design without replacement is a probability function $P(S)$ satisfying: $\sum_{S\in E} P(S) = 1$ and $P(S) \geq 0$ for all $S \in E$. Where E is the set of all the samples of n distinct units from U.

A common problem in survey sampling is to estimate the total

$$T = \sum_{i=1}^{N} Y_i$$

of an unknown positive variable measured without error. Y_i being the value of this variable for the i-th unit of U. One of the best strategies is to use the Horvitz-Thompson (1951) estimator

$$\hat{T} = \sum_{i\in S} Y_i / \pi_i ,$$

where π_i's are the first-order inclusion probabilities of the sampling design $P(S)$. In this paper, we analyse a πps sampling design with π_i proportional to the strictly positive value X_i of a known auxiliary variable for the i-th unit. We assume that $nX_i \leq \sum_{i=1}^{k} X_i$ for all i and k such that $i \leq k$ and $n < k \leq N$.

The variance of $\hat{T}$ is given by the Yates-Grungy (1953) estimator

$$V(\hat{T}) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j<i}}^{N} (\pi_i \pi_j - \pi_{ij}) \left[(Y_i \pi_i^{-1}) - (Y_j \pi_j^{-1}) \right]^2 .$$

π_{ij} is the second-order inclusion probability of the units i and j. If $\pi_{ij} > 0$ for all i and j then an unbiased estimator of $V(\hat{T})$ is given by

$$\hat{V}(\hat{T}) = \sum_{i\in S} \sum_{\substack{j\in S \\ j<i}} (\pi_i \pi_j \pi_{ij}^{-1} - 1) \left[(Y_i \pi_i^{-1}) - (Y_j \pi_j^{-1}) \right]^2 .$$

The double sum feature of this variance is inconvenient. Moreover, the variance estimator requires heavy calculation because of the second-order inclusion probabilities. The aim of this paper is to give an acceptable approximation of the variance that does not need the second order-inclusion probabilities.

[1] Yves G. Berger, Université Libre de Bruxelles, 44, Av. Jeanne, CP124, LB190, B-1050 Brussels, Belgium; e-mail: yvberger@ulb.ac.be.

Our approach is asymptotic in a sense that

$$d = \sum_{i=1}^{N} \pi_i (1 - \pi_i) \to \infty.$$

This assumption implies that n and N-n tends to infinity.

The symbol $O(\varepsilon)$ will denote a number such that $O(\varepsilon)/\varepsilon$ remains bounded if ε tends to zero. $o(1)$ will be a number such that $o(1)$ tends to zero if d goes to infinity.

3. THE CHAO SAMPLING

The Chao sampling (1982) is implemented by selecting the n first units of U with probability 1. Each unit $i\,(i = n+1, ..., N)$ is selected with probability $nX_i / \sum_{j=1}^{i} X_j$. Each time a unit i is selected, a unit drawn at random within the units already selected is replaced by unit i.

Chao's sampling is a without replacement *πps* sampling design with fixed sample size. It is a sequential sampling in the sense that the sample is drawn in one pass through the population. This sampling design is a generalization of the McLeod and Bellhouse (1983) sampling design. The Chao sampling is sequential as the items are selected through a sequential run of the population. Furthermore, the second-order inclusion probabilities are always are strictly positive. Thus the Yates-Grungy estimator is unbiased.

4. VARIANCE ESTIMATOR USING CHAO'S SAMPLING

Berger (1997a) shows that if $i < j$,

$$\pi_i \pi_j \pi_{ij}^{-1} - 1 = \begin{cases} \beta_{ij} & \text{if } j \leq n+1, \\ \alpha_j & \text{if } j > n+1, \end{cases}$$

where

$$\beta_{ij} = -1 + X_i\, X_j\, C_n^{-1} (X_i\, X_j - C_n n^{-1})^{-1} p_{n+1},$$

$$\alpha_j = -1 + n(n-1)^{-1} (1 - X_{j-1} C_{j-1}^{-1})\, p_j,$$

$$p_i = \prod_{k=i}^{N} (1 - X_k C_k^{-1})^2 (1 - 2 X_k C_k^{-1})^{-1},$$

$$C_i = \sum_{k=1}^{i} X_k.$$

Thus only the α_j's and β_{ij}'s are required to compute the Yates-Grungy variance estimator. Finally, approximately n quantities need to be computed instead of $n(n-1)/2$ second-order inclusion probabilities.

5. APPROXIMATION OF THE VARIANCES

A first attempt to approximate the variance estimator can be found in Berger (1996). This paper proposes an approximation of the second-order inclusion probabilities of the Chao sampling

$$\tilde{\pi}_{ij} = \pi_i\, \pi_j (n-1)(n - \pi_j)^{-1} \text{ if } i < j.$$

Replacing π_{ij} by $\tilde{\pi}_{ij}$ in $\hat{V}(\hat{T})$, we have an approximation $\tilde{\sigma}^2$ of the variance estimator. Berger (1997d) shows that $\tilde{\sigma}^2$ is close to $\hat{\sigma}^2$, in the sense that

$$|(\tilde{\sigma}^2 / \hat{\sigma}^2) - 1| \leq \left[\pi_{(N)} - \pi_{(1)}\right]\left[1 - \pi_{(N)}\right]^{-1},$$

where $\pi_{(N)}$ and $\pi_{(1)}$ are respectively the greater and the smaller first-order inclusion probabilities. $\hat{\sigma}^2$ is defined by

$$\hat{\sigma}^2 = n(n-1)^{-1} \hat{d} d^{-1} \left[\sum_{i \in S} y_i^2 (1 - \pi_i) \pi_i^{-2} - \hat{d} \hat{G}^2\right],$$

with

$$\hat{G} = \hat{d}^{-1} \sum_{i \in S} Y_i (1 - \pi_i) \pi_i^{-1},$$

$$\hat{d} = \sum_{i \in S} (1 - \pi_i).$$

We note that $\hat{\sigma}^2$ does not depend on the second-order inclusion probabilities. Moreover, this estimator is as simple as the Hansen-Hurwitz (1943) estimator. $\hat{\sigma}^2$ is an estimator of the Hájek variance σ^2 defined by

$$\sigma^2 = N(N-1)^{-1} \left[\sum_{i=1}^{N} Y_i^2 (1 - \pi_i) \pi_i^{-1} - dG^2\right],$$

with

$$G = d^{-1} \sum_{i=1}^{N} Y_i (1 - \pi_i).$$

It is possible to have a rate of convergence of the approximations σ^2 and $\hat{\sigma}^2$. Berger (1998b) shows that

$$\left|\left[V(\hat{T}) / \sigma^2\right] - 1\right| \leq O(D^{1/2}) + |o(1)|,$$

$$\left|\left[\hat{V}(\hat{T}) / \hat{\sigma}^2\right] - 1\right| \leq O(D^{1/2}) + |o(1)|,$$

where

$$D = \sum_{S \in E} P(S) \log\left[P(S) / R(S)\right]$$

is the divergence of the sampling $P(S)$. Thus if D is small, σ^2 and $\hat{\sigma}^2$ are valid approximations. Berger (1997d) shows that if we implement the Chao sampling then

$$D<-\frac{1}{2}\log\left[1-n(1+V_{\pi}^{2})N^{-1}\right]+n\varepsilon_1\left[\varepsilon_1+\varepsilon_2-\log(\varepsilon_1)\right]$$

$$-\varepsilon_1\left[\frac{1}{2}\log(2\pi n)-\pi_{n+2}\right]+\log\left[1+o(1)\right]+n^{-1}\beta(1-\varepsilon_1),$$

where

$$\varepsilon_1 = C_{n+1}C_N^{-1},$$

$$\varepsilon_2 = \mathrm{Max}\,\{nX_iC_{n+1}^{-1}-\log(nX_iC_{n+1}^{-1})-1 : i=1,\ldots n+1\}.$$

β is a constant such that $0<\beta<1/12$. V_π is the coefficient of variation of the first-order inclusion probabilities. As ε_2 is bounded, D goes to zero if n/N and $n\varepsilon_1$ tend to zero.

For example, if $n = 10, N = 1{,}000, V_\pi = 1, \varepsilon_1 = 0.0001$ and $\varepsilon_2 = 0.01$ then the ratio of the standard deviation lies between 0.88 and 1.11. In this case, σ^2 and $\hat{\sigma}^2$ approximate excellently the target standard deviations.

6. ASYMPTOTIC NORMALITY

If D goes to zero then the Horvitz-Thompson estimator has an asymptotic normal distribution [see Berger (1998a)]; *i.e.*,

$$|\mathrm{pr}\left[(T-\hat{T})/\sigma^2 \le y\right]-\Phi(y)| \to 0$$

if e and D go to zero and d tends to infinity. Where

$$e = \mathrm{Min}\,\{z : \sigma^{-2}\sum_{i:|Y_i-G\pi_i|>z\pi_i\sigma}(Y_i-G\pi_i)^2(1-\pi_i)\pi_i^{-1}\le z\}.$$

$\Phi(y)$ defines the normed normal distribution function.

Furthermore, if there are positive constants b_1 and b_2 such that

$$\sum_{i=1}^{N}(Y_i/\pi_i)^4 < b_1N \text{ and } \sigma^2 = b_2N,$$

then

$$|\mathrm{pr}\left[(T-\hat{T})/\sigma^2 \le y\right]-\Phi(y)| < O(N^{-1/2})+O(D^{1/2}).$$

This gives a rate of convergence to the normal distribution. Finally, if n/N and $n\varepsilon_1$ tend to zero then the Horvitz-Thompson estimator has a normal distribution.

7. CONCLUSION

If the sample is selected with the Chao sampling, we can compute the exact value of the Yates-Grungy estimator without the second-order inclusion probabilities. Under special conditions, the variance can be approximated by $\hat{\sigma}^2$ and the Horvitz-Thompson estimator has a normal distribution.

REFERENCES

Berger, Y.G. (1996). Asymptotic variance for sequential sampling without replacement with unequal probabilities. *Survey Methodology*, 22, 167-173.

Berger, Y.G. (1997a). Variance estimation using list sequential scheme for unequal probability sampling. To appear in *Journal of Official Statistics*, (Statistics Sweden).

Berger, Y.G. (1997d). Inference using a sequential sampling with unequal probabilities. Preprint de *l'Institut de Statistique et de Recherche Opérationnelle de l'ULB*.

Berger, Y.G. (1998a). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 67, 209-226.

Berger, Y.G. (1998b). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. To appear in the *Journal of Statistical Planning and Inference*.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annal of Mathematical Statistics*, 35, 1491-1523.

Hájek, J. (1981). Sampling from a Finite Population. *Marcel Dekker, Inc.*, New York and Bassel.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite population. *Annal of Mathematical Statistics*, 14, 333-362.

Horvitz, D.G., and Thompson D.J. (1951). A Generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

McLeod, A.I., and Bellhouse, D.R. (1983). A convenient algorithm for drawing a simple random sample. *Applied Statistics*, 32, 2.

Madow, W.G. (1949). On the theory of systematic sampling II. *Annal of Mathematical Statistics*, 20, 333-354.

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Serie B, 1, 253-261.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

SAMPLE DESIGN FOR THE PRODCOM (PRODUCT SALES) INQUIRY

R. Chambers, M. Cruddas and P. Smith[1]

ABSTRACT

The PRODCOM (*Prod*ucts of the European *Com*munity) inquiry is a EC-wide business survey designed to produce country level estimates of product sales for a large range of EC manufactured products. In the UK the survey is used to produce estimates of value of total production by product. Design of the UK version of the survey presents an interesting challenge, in that a design is required which balances the fact that most businesses are only involved in a small number of products against the need for accurate estimates for a wide range of products. This paper presents a potential design methodology for this situation based on minimising the average relative variance of the range of product estimates required given a model for product sales which explicitly allows for the fact that most businesses are only involved in producing a small range of products. Numerical results illustrating the expected accuracy of the product estimates obtained under this design are also presented.

KEY WORDS: Sample design; Multivariate allocation; Product surveys.

1. INTRODUCTION

PRODCOM stands for PRODucts of the European COMmunity. The UK PRODCOM Inquiry collects values (sales invoiced) and volumes (units and/or weights) of product data from UK manufacturers for the most commonly manufactured products within the European Community. It is part of a Europe-wide harmonised system for the collection and publication of product statistics carried out by EUROSTAT (the Statistical Office of the European Community).

The PRODCOM inquiry asks each of the units it approaches for information about the products in the PRODCOM product list. There are approximately 4,800 products in the list, so in effect there are 4,800 outcome variables. Since companies contribute to only a limited number of the products there will be an overwhelming number of zero responses, either actual or implied, for each of these outcome variables, making the use of the standard ratio estimator (Cochran 1977) inappropriate. Rather than ask for information on each of the 4,800 products in the list, the PRODCOM inquiry sends personalised questionnaires to the sampled businesses covering products that they are known to produce and products that are thought to be within their scope following consultation with trade associations and businesses.

The EUROSTAT requirement is to collect product information from businesses that account for 90% of total sales in the manufacturing sector (European Commission 1991), but the UK survey aims to produce estimates for the total sales of each commodity. Since total sales is unknown, employment, available for all businesses on the Inter-Departmental Business Register (IDBR) (Perry 1995), is used as a proxy measure of coverage. Selection of the sample is made from the IDBR. Sampling is stratified, with strata defined by 5-digit industries of the Standard Industrial Classification 1992 (SIC92) (CSO 1992), and the size of the businesses measured by their employment using information from the IDBR for each business in the population. PRODCOM has two periodicities depending on the industry being sampled; we will concentrate on the annual PRODCOM inquiry which samples 24,000 businesses in 200 industries.

The task in hand is to produce a design which accurately estimates the total sales for each commodity, for a fixed number of survey forms of equal cost. This paper provides a general framework for an alternative to the usual ratio estimator that uses the zero returns in the sample to predict zero returns in the non-sampled units in order to produce estimates of product totals. The model is then used to provide a two-stage method of allocating a sample of fixed size to industry by size strata based on minimising the average variances of the individual product totals adjusted for the size of the totals (the average relative variance).

2. A MODEL FOR PRODCOM

The ratio estimation model makes sense if the population values of Y are strictly positive. However, the presence of large numbers of zeros for many PRODCOM outcome variables makes the use of this model invalid. The new model proposes that the probability of a non-zero value for an outcome variable is a function of size (larger units are unlikely to have the same zero response propensity as small ones) and that the size of non-zero responses increases with the size of the auxiliary variable.

[1] Ray Chambers, University of Southampton, UK; Marie Cruddas and Paul Smith, Office for National Statistics, UK, Cardiff Road, Newport, NP9 1XG, UK, e-mail: paul.smith@ons.gov.uk.

This can be written:

$$Pr(Y_i > 0 | X_i) = \pi(X_i)$$

$$E(Y_i | X_i, Y_i > 0) = \beta X_i$$

$$\mathrm{Var}(Y_i | X_i, Y_i > 0) = \sigma^2 v(X_i)$$

Now the parameters β and σ^2 can be estimated by the usual ratio estimators over the units in the sample with non-zero response.

$$\hat{\beta} = \frac{\sum_s \frac{\Delta_i Y_i X_i}{v(X_i)}}{\sum_s \frac{\Delta_i X_i^2}{v(X_i)}},$$

$$\hat{\sigma}^2 = \frac{\sum_s \frac{\Delta_i (Y_i - \hat{\beta} X_i)^2}{v(X_i)}}{\sum_s \Delta_i - 1}$$

where Δ_i denotes whether the sampled unit has a non zero response or not, that is

$$\Delta_i = \begin{cases} 1 & \text{if } Y_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

This model implies that $E(Y_i | X_i) = \pi(X_i)\,\beta X_i$, so an estimate of the total is

$$\hat{T} = \sum_s Y_i + \hat{\beta} \sum_r \hat{\pi}(X_i) X_i$$

where $\hat{\pi}(X_i)$ denotes an appropriate estimate of $\pi(X_i)$. It can be shown (see Chambers, Cruddas & Smith (forthcoming)) that, under this model, the variance of the sample error $T - \hat{T}$ is approximately

$$\mathrm{Var}(\hat{T} - T) \approx \sigma^2 \left[\frac{\left(\sum_r E(\hat{\pi}(X_i))\, X_i \right)^2}{\sum_s \pi(X_i) X_i^2 / v(X_i)} + \sum_r \pi(X_i) v(X_i) \right] +$$

$$\left[\beta^2 + \frac{\sigma^2}{\sum_s \pi(X_i) X_i^2 / v(X_i)} \right] \mathrm{Var}\left(\sum_r \hat{\pi}(X_i) X_i \right) +$$

$$\beta^2 \sum_r \pi(X_i)(1 - \pi(X_i))\, X_i^2$$

In order to use this model the functions $\pi(X_i)$ and $v(X_i)$ must be specified, as well as an estimator of $\pi(X_i)$. A natural formulation for the $\pi(X_i)$ is to make it constant within strata (denoted by π_h where h is the stratum indicator), allowing it to differ between strata. An unbiased estimator for π_h is then the proportion of units in stratum h with non-zero values of the outcome variable.

Two special cases for the $v(X_i)$ term were considered:

Model 1: $v(X_i) = X_i$, leading to estimated total $\hat{T}_1$.

Model 2: $v(X_i) = X_i^2$, leading to estimated total $\hat{T}_2$.

3. SAMPLE ALLOCATIONS

The problem of allocating a sample of fixed size to the industry by size strata is complicated by the large number of outcome variables. The approach taken here is to first allocate the sample to industry strata ignoring the size strata, using an allocation which minimises the average relative variance of the individual product totals across the industry strata. Then once these industry level samples have been set, allocation of the sample to size strata within each industry stratum is carried out using a similar minimum average relative variance criterion within industry stratum.

Note also that the variance formulae used in the allocation process have been simplified by only including the two leading terms in each variance expression, to avoid the computational difficulty of producing sums of squares for the population values.

3.1 Sample Size Allocation Across Industry Strata

For both of the models it can be shown that:

$$\mathrm{Var}(\hat{T}_m(a,c) - T(a,c)) \propto \frac{K_m(a,c)}{n}$$

where $T(a,c)$ is the total production of commodity c by industry a, m denotes either model 1 or model 2, and

$$K(a,c) = \frac{\sigma^2 \left(\sum_h \pi_h F_h \bar{X}_h \right)^2}{\sum_h \pi_h f_h D_h} +$$

$$\beta^2 \sum_h \pi_h (1 - \pi_h) F_h^2 f_h^{-1} \bar{X}_h^2$$

$$\text{where } D_h = \begin{cases} \bar{X}_h \text{ if } & v(X) = X \\ 1 \text{ if } & v(X) = X^2 \end{cases}$$

The variances for different commodities may differ wildly because of the different levels of commodity values (within an industry the totals of mainline commodities are far higher than other commodities). Thus it is appropriate to minimise the average of the relative variances rather than of the variances themselves.

Allocating a sample of size n_a, from the total sample n, to industry a on the basis of minimising:

$$\sum_c \frac{1}{T^2(c)}(\mathrm{Var}(\hat{T}(c) - T(c))$$

where $T(c)$ is total production of commodity c across all industries, leads to

$$n_a = n \frac{\left[\sum_c \frac{1}{T(c)^2} K(a,c)\right]^{½}}{\sum_{a^*}\left[\sum_c \frac{1}{T(c)^2} K(a^*,c)\right]^{½}}$$

However the $T(c)$ must then be estimated and this will depend on the allocation. An approximation to the $T(c)$ is to use the values of $T(c)$ achieved under the current allocation, and that is what is done here. Note that each of the two models yields a different allocation, so we average over these allocations to give a single final allocation.

3.2 Optimal Size Stratum Allocation

Once the industry level allocation has been set the next step is to allocate to the size strata. The approach is again to minimise the average relative variances of the products that occur within a particular industry.

For any commodity c in an industry stratum, the equations in section 2 for $\mathrm{Var}(\hat{T}(c) - T(c))$ can be written as:

$$\mathrm{Var}(\hat{T}(c) - T(c)) \approx$$

$$\frac{N^2}{n}\left[\sigma_c^2 \frac{(\sum_h \pi_{hc} F_h \bar{X}_h)^2}{\sum_h \pi_{hc} f_h \bar{D}_{sh}} + \beta_c^2 \sum_h \pi_{hc}(1-\pi_{hc}) F_h^2 f_h^{-1} \bar{X}_h\right]$$

where

$$f_h = \frac{n_h}{n}, \text{ the fraction of the sample in size stratum } h$$

$$F_h = \frac{N_h}{N}, \text{ the fraction of the population in size stratum } h$$

$$\bar{X}_{sh} = \frac{1}{n_h}\sum_{s_h} X_i$$

$$\bar{X}_h = \frac{1}{N_h}\sum_{rh} X_i$$

and

$$D_h = \begin{cases} \bar{X}_{sh} & \text{if } \quad v(X) = X \\ 1 & \text{if } \quad v(X) = X^2 \end{cases}$$

Hence the relative variance for an industry can be written

$$RV = \frac{N^2}{n}\sum_c \frac{1}{T^2(c)}\left[\frac{A_c}{\sum_h \alpha_{hc} f_h} - B_c \sum_h \beta_{hc} f_h^{-1}\right]$$

where

$$A_c = s_c^2\left(\sum_{h=1}^{H} p_{hc} F_h \bar{X}_h\right)^2$$

$$B_c = \beta_c^2$$

$$\alpha_{hc} = p_{hc} \bar{D}_{sh}$$

$$\beta_{hc} = p_{hc}(1 - p_{hc}) F_h^2 \bar{X}_h^2$$

Note that the term N^2/n does not affect the optimisation and will be dropped from here on.

We want to find optimal sample stratum allocation proportions $\{f_h\}$ that minimise the RV above, such that $\sum_h f_h = 1$ and $0 \le f_h \le 1$. The first term in the RV is minimised on the boundary. That is, there is a stratum h^* such that the first term is minimised when all of the sample is in h^* (that is $f_h = 1$ when $h = h^*$ and zero otherwise). This means that h^* is the stratum with the minimum value of

$$\sum_c \frac{A_c}{T^2(c)\alpha_{hc}}.$$

The second part of the RV conforms to the standard methods for finding a minimum subject to the constraints. Put

$$\beta_h^* = \sum_c \frac{B_c \beta_{hc}}{T^2(c)}.$$

The values of f_h that minimise the second component of the RV are then given by

$$\hat{f}_h = \frac{\beta_h^{*-½}}{\sum_{g=1}^{H} \beta_g^{*-½}}.$$

Thus finding the allocation that minimises the RV amounts to finding h^*, optimising the second part of the RV then progressively adding small amounts to f_h^*, fixing it and minimising the second part until the RV starts to get larger. Hopefully this will happen before the boundary is reached.

3.3 Strategy for Finding an Optimal Size Stratum allocation

The final algorithm for the allocation of sample to the size strata within an industry is then

1. Find the stratum h^* as indicated above.
2. Compute the initial allocation.

3. Calculate the RV for this allocation, RV(0).
4. Modify the initial allocation by adding a small positive amount to f_h^* and recalculate the values of the other f_h's.
5. Calculate the RV for this allocation, RV(1).
6. If the new allocation gives a smaller relative variance than the initial allocation make the new allocation the initial allocation and the new RV(0) = RV(1) and go back to step 4. Otherwise take the initial allocation as optimal and stop.

For practical application, some other constraints have been included in the sampling scheme, in order to comply with policy for ONS business surveys. The main ones are:

- businesses with 0-9 employment have been allocated separately with their own fixed sample size to ensure that the number sampled is small;
- sampling fractions $\in [0.75, 1.0]$ are rounded to 1.0, and fractions $\in (0.5, 0.75)$ are rounded to 0.5, in both cases to avoid strange rotation patterns in rotational sampling;
- a minimum sample size in each industry was applied to ensure that result for each industry would be based on a reasonable number of observations.

4. RESULTS

Firstly separate allocations were produced using variances calculated from previous survey data and using models 1 and 2. These allocations were then averaged to give an overall allocation. This was done twice, once using 4-digit products and once using 8-digit products. The expected sampling relative standard errors from these new allocations were calculated for estimation according to models 1 and 2, and these are shown in Table 1.

Note that the current design is subject to some non-response; imputations have not been included when estimating variances, so that the effective sample size is smaller than the nominal 24,500, and the achieved allocation may be suboptimal due to the non-response process. In contrast the rse's from the new allocation assume full response, and hence provide lower bounds for those expected in practice.

From these results it can be seen that the new method and 24,500 forms results in lower mean and median rse's when compared with the current sample size. This gain in accuracy is kept at the smaller sample size of 20,000 when allocations are based on 4-digit product groups, but is not the same when 8-digit groups are used. In order to maintain the accuracy of product information from the surveys, and based on these results, the PRODCOM sample size will be maintained at 24,500, but with the new allocation.

There are several areas within the model which will benefit from further investigation, and it is hoped that the current design can be kept stable for a few years to accumulate data to test these hypotheses, and eventually to feed into a further redesign which will again enable the accuracy of the sampling to be increased.

REFERENCES

Chambers, R., Cruddas, M., and Smith, P.A. (forthcoming) Sample design for the PRODCOM Inquiry.

Cochran, W. G. (1977) *Sampling Techniques*. New York: Wiley.

CSO (1992). *Standard Industrial Classification of economic activities 1992*. Newport: HMSO.

European Commission (1991). Council Regulation (EEC) No 3924/91 of 19 December 1991.

Perry, J. (1995). The Inter-Departmental Business Register. *Economic Trends*, 505, November 1995.

Table 1

The mean, median and maximum relative standard errors over product groups at 4-digit and 8-digit levels of the product classification using three scenarios: (a) the current design; (b) a reallocation using the methods of this paper with overall sample size of 24,500; and (c) a reallocation with overall sample size 20,000

4-digit product group variances						
	(a) current		(b) n = 24,500		(c) n = 20,000	
	model 1	model 2	model 1	model 2	model 1	model 2
mean rse	4.3	5.0	3.9	4.6	4.1	4.8
median rse	1.8	2.0	1.5	1.7	1.7	1.9
maximum rse	67.1	113.2	91.1	153.6	91.1	153.6
8-digit product group variances						
	(a) current		(b) n = 24,500		(c) n = 20,000	
	model 1	model 2	model 1	model 2	model 1	model 2
mean rse	11.1	11.8	10.2	10.9	11.5	12.3
median rse	6.5	7.0	6.2	6.7	6.9	7.5
maximum rse	258.9	103.8	207.7	128.8	238.3	151.0

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

TELEPHONE INTERVIEW SAMPLE DESIGNS AT THE U.S. CENSUS BUREAU

D. Garrett, J. Hartman and A. Meier[1]

ABSTRACT

Sample designs from two maximized telephone surveys are described. Research on data from these surveys indicates that households not interviewed by telephone centres are different from households which are interviewed by telephone centres. A weighting adjustment for telephone centre nonresponse is described and evaluated. An RDD sample is compared to a full coverage sample.

KEY WORDS: CATI; RDD; Nonresponse; Weighting.

1. INTRODUCTION

With data collection costs increasing and survey budgets decreasing, the Census Bureau wants to explore alternative methods to maximize the use of telephone interviewing. Two recent surveys conducted by the Bureau used alternative telephone interview sample designs.

The 1995 American Travel Survey (ATS) sample had two parts. One part used field interviewers to collect data. This method of interviewing will be referred to as decentralized Computer Assisted Interviewing (d-CAI). The other part was interviewed solely by centralized Computer Assisted Telephone Interviewing (c-CATI) and non-interviews were not followed up.

The 1996 FHWAR (Survey of fishing, hunting, and wildlife association recreation) had two samples. One sample was interviewed using our usual c-CATI methods with a subsample of non-interviews followed up by field interviewers. The other sample used Random Digit Dialling (RDD) and was selected for research purposes. Non-interviews were not followed up.

2. AMERICAN TRAVEL SURVEY (ATS)

2.1 Background and Sample Design

The main purpose of the ATS was to provide estimates of the number of trips 100 miles or more (one-way) from home, and other travel-related items for those trips, for the 1995 calendar year.

The 1995 ATS used approximately 80,000 residential units from the 1980 state-based Current Population Survey (CPS) last interviewed between December 1990 and October 1994. In addition to the address of each sample unit, we also had the phone number for a large proportion of the CPS sample.

After selecting the initial sample, we used several sources to try to get a good phone number. First, we sent the sample addresses to a telephone research firm called Telematch. They provided phone numbers for about 35 percent of the cases but most of the phone numbers (90%) were the same as the one we had from CPS.

In January of 1995, we tried calling each number we had to verify the phone number for the sample unit. If the interviewer couldn't reach the address with the existing phone number, we used a telephone look-up operation to try to find the correct phone number using sources such as the apartment manager, the tax assessor, the utility company, post offices, telephone number CD-ROM, and the 911 administrator.

We confirmed about 55% of the phone numbers. Since there still weren't enough cases for the c-CATI sample, we added some extra CPS sample.

2.2 Splitting the Sample

We now had two types of units: units with confirmed (*i.e.*, good) phone numbers and units with unconfirmed (*i.e.*, bad) phone numbers. We split the units with confirmed phone numbers between the Complete Coverage and Good Phone Frames. We selected a subsample of units with unconfirmed phone numbers and put them in the Complete Coverage Frame.

The Complete Coverage Frame had 23,201 units with confirmed phone numbers and 11,981 with unconfirmed phone numbers and represented all living quarters in the U.S. Units in this frame were interviewed by both decentralized CATI (d-CATI) and Computer-Assisted Personal Interviewing (CAPI). (The two methods of d-CATI and CAPI are together referred to as d-CAI, since they are both decentralized Computer-Assisted Interviewing.) A field interviewer recorded the results of the interview on a laptop computer. The field interviewers conducted about 88% of all the interviews by phone in order to maximize the telephone interviewing. For the cases with confirmed phone numbers over 90% of the interviews were by telephone.

[1] Dale Garrett, U.S. Department of Commerce, Bureau of the Census, Washington, D.C. 20233, U.S.A.; e-mail: dgarrett@census.gov.

The Good Phone Frame had 44,983 units with confirmed phone numbers. It didn't represent all living quarters in the U.S. Units in this frame were interviewed by c-CATI. Non-interviews weren't followed up by personal visit. They remained as non-interviews and were represented by comparable units in the Complete Coverage frame in a weighting adjustment.

2.3 Weighting Adjustment for c-CATI Nonresponse

The weighting procedure adjusted for c-CATI non-interviews from the Good Phone Frame. Since we couldn't always identify the non-interview reason for the Good Phone Frame units, we grouped all non-interviews together. The following types of c-CATI non-interviews were included in the adjustment:

- units that refused to be interviewed;
- units that no longer exist or are used for non-residential purposes;
- units that no longer have phones or have unlisted phone numbers;
- units that use an answering machine to screen calls;
- units whose new phone number isn't available;
- vacant units.

To represent the Good Phone Frame non-interviews in the weighting adjustment we identified comparable units with confirmed phone numbers. These included d-CAI non-interviews plus all d-CAI interviews which we thought that c-CATI would not have interviewed. To simplify we refer to these interviews as d-CAI cases which c-CATI could not interview. The remainder of d-CAI interviews with confirmed phone numbers are referred to as d-CAI cases which c-CATI could interview.

For interviews in the Complete Coverage Frame, we first looked to see if c-CATI had access (in theory) to the phone number. If the number provided by the respondent was different from the one we gave the interviewer, we sent the unit to a telephone number look-up operation. If the look-up operation didn't find the new number, we included the case in the c-CATI could not interview group.

If the respondent provided the same number we gave the interviewer or the look-up operation found the number, we looked at information collected during the interview. The information sought to identify the following conditions which indicate c-CATI would not have been able to interview the unit and that we should include it in the c-CATI could not interview group for the weighting adjustment.

- the respondent requested a personal visit;
- the respondent screens calls on his or her answering machine;
- the interview was conducted in a language unavailable at the phone centres;
- the interviewer tried to contact the respondent several times by phone but was unable to (*e.g.*, they got a ring-no-answer);
- the interviewer notes indicate c-CATI wouldn't have been able to conduct the interview (*e.g.*, respondent was hard of hearing).

2.4 Response Rates for the 1995 American Travel Survey

Looking at the confirmed phone number cases the d-CAI response rate was much better than c-CATI. The d-CAI response rate was 87% with 9% eligible non-interviews and 4% ineligibles. The c-CATI response rate was 77% with an estimated 19% eligible non-interviews. An eligible non-interview is a unit considered occupied residential but for which data was not collected because no one was home, respondent refused or otherwise could not be interviewed.

2.5 Research Related to the ATS Sample Design

The ATS design allows us to evaluate the relative effectiveness of c-CATI interviewing without follow-up of non-interviews. The research should help us to determine how different the cases which c-CATI could interview are compared to cases which could not be interviewed by c-CATI. It will also evaluate our weighting adjustment for c-CATI nonresponse. The next ATS sample design and weighting procedures will be developed using the results of this research.

Here are some of the comparisons planned for ATS research. Many of these comparisons have already been done for demographic and/or travel characteristics using unbiased weights:

1. c-CATI non-interviews *vs* the d-CAI cases classified as c-CATI could not interview from the Confirmed Phone Numbers (using CPS data);
2. c-CATI non-interviews *vs* c-CATI interviews (using CPS data);
3. d-CAI cases classified as c-CATI could not interview *vs* c-CATI could interview (from the Confirmed Phone Numbers);
4. c-CATI interviews *vs* d-CAI Confirmed Phone Number households;
5. c-CATI interviews *vs* d-CAI All Households with Telephones;
6. c-CATI interviews *vs* d-CAI All Households;
7. households with phones vs households without phones;
8. small subsample of c-CATI non-interviews *vs* c-CATI interviews;
9. additional research on the c-CATI nonresponse adjustment for the Good Phone Frame.

The first three comparisons help us to evaluate our ATS weighting adjustment for c-CATI nonresponse. If the weighting adjustment was effective, the c-CATI nonrespondents should be similar to the units used in the weighting to represent them for characteristics important to the survey. Also the c-CATI nonrespondents should have differences from the c-CATI respondents for such characteristics otherwise the weighting adjustment is not needed. However, we cannot compare the c-CATI nonrespondents directly since we don't have ATS data for them. Instead the first two comparisons use older data collected by CPS for comparisons so that we have some data for the units which became the current c-CATI nonrespondents. The third comparison relies on our classification of d-CAI interviews to compare the cases c-CATI could and could not interview.

The starting point in evaluating our weighting adjustment for c-CATI nonresponse was to look at comparisons 1-3. If comparison 1 showed mostly small differences that were not significant and comparisons 2 and 3 showed many large and often significant differences for the ATS and CPS characteristics, that should provide a confirmation that our weighting adjustment had a positive impact.

The fourth, fifth and sixth comparisons will indicate how well ATS demographic and travel characteristics collected by c-CATI interviews from the Good Phone Frame can be used to represent part or all of the universe.

The seventh comparison highlights the differences between households which have phones and those which don't. The eighth comparison uses a few items collected by personal visit to help clarify the differences between c-CATI respondents and nonrespondents. The ninth will provide an alternative evaluation of our c-CATI nonresponse adjustment by reweighting before doing the comparisons. No results are available for 8 and 9.

In general, if we tend to find large significant differences for Comparisons 2-8, this suggests that it is important to include cases which c-CATI could not interview in our estimates.

2.6 Preliminary Results of the ATS Research for Comparisons 1 Through 7

Comparisons 1, 2 and 3: These comparisons together seem to support the need for the ATS weighting adjustment for c-CATI nonresponse using d-CAI interviews. Comparison 1 showed small differences which tended not to be significant for the following CPS items: sex, race and education of reference person, owner-occupied unit and income. Comparison 2 showed large and significant differences for the same items. For comparison 3, Table 1 shows many demographic differences which were significant at the national level. Concerning trip characteristics, most were not significant at the national level, however each characteristic tested showed between 10 to 18 states with significant differences. Comparison 9 will be done later to complete the analysis of the weighting adjustment.

Comparisons 3-7 which used only ATS data are summarized in Table 1.

Comparisons 4, 5 and 6: The c-CATI interviews from the Good Phone Frame were compared to three groups. In comparison 4 we did find some significant differences with the Confirmed Phone Number Households as we expected since the c-CATI non-interviews were not followed up. The all telephone comparison includes telephone households from the unconfirmed phone group so had several differences from the Good Phone Frame. However, the most outstanding differences were for Comparison 6, since non-telephone households were included in the comparison of c-CATI to the Complete Coverage Frame. Together these comparisons show the importance of the Complete Coverage Frame to complement the Good Phone Frame design used by ATS. (See Table 1)

Comparison 7: About 6% of the USA households don't have telephones in the house or apartment. As expected, they had large significant differences for many demographic characteristics, proportion of travelling households and some trip characteristics. (See Table 1).

Table 1
Summary of Significant Differences for ATS
Comparisons 3-7

Characteristic	3	4	5	6	7
Proportion Travelling Households	Yes	Yes	Yes	Yes	Yes
Spouse Present (Couples)	Yes	No	Yes	Yes	Yes
Owner Occupied	Yes	Yes	Yes	Yes	Yes
Full-time Job	States	States	No	No	Yes
White Reference Person	Yes	No	Yes	Yes	Yes
Black Reference Person	Yes	No	Yes	Yes	Yes
Hispanic Reference Person	Yes	Yes	Yes	Yes	Yes
Elderly Respondents	Yes	No	Yes	Yes	Yes
Single with Small Children	Yes	Yes	Yes	Yes	Yes
High School Graduate	Yes	Yes	Yes	Yes	Yes
Lower Income/ Higher Income	Yes	Yes	Yes	Yes	Yes
Trips by Automobile	States	No	States	States	Yes
Trips by Airplane	States	No	No	No	Yes
Business Trips	States	No	No	States	States
Pleasure Trips	States	No	No	No	States
Trips with less than 400 Miles Round Trip Distance	States	No	No	No	Yes

Note: Yes - significant at $\alpha = .01$ for U.S.A.
States - not significant for U.S.A. but significant at $\alpha = .1$ for 8 or more states
No - other

General: The significant and sometimes large differences seen in these comparisons help us to see how telephone households are different from non-telephone households and how c-CATI respondents are often different from c-CATI nonrespondents as well as how the Good Phone Frame estimates are often different from the complete universe. The research will help us to reduce potential bias during the design of the next ATS by finding altemate ways to incorporate aspects of a full coverage frame into a maximized CATI design.

3. SURVEY OF FISHING, HUNTING, AND WILDLIFE ASSOCIATED RECREATION (FHWAR)

A study was conducted concurrently with the 1996 FHWAR to identify and measure bias in using an RDD methodology. The regular FHWAR sample of 77,000 households was originally based on expired sample from the 1980 state-based CPS which had been updated by new construction sample. In order to maximize telephone interviewing, telephone numbers were updated by research for many of the addresses. The sample was interviewed first in centralized CATI for thirty days. Then a subsample of those cases which couldn't be contacted in c-CATI were sent to the field for follow-up interviews by d-CAI. Recall that d-CAI includes both decentralized CATI and CAPI interviewing. We completed screening interviews for about 44,000 households and completed detailed sportsmen questionnaires for about 23,000 participants.

For the RDD study, we randomly selected 23,000 telephone numbers for the national comparison. The households identified by these numbers were interviewed from the centralized CATI facility with no field follow-up. We completed screening interviews for about 7,000 households and completed detailed sportsmen questionnaires for about 3,000 participants.

3.1 Research Plans for FHWAR/RDD

The primary purpose of the study is to measure the bias in the detail estimates from the RDD survey. Additionally, we will compare selected estimates from subsets of the full survey to see how much we gain by using CAPI recycling. We will produce three sets of estimates – the FHWAR estimates, the RDD estimates, and FHWAR estimates using only cases which were completed with centralized CATI.

In addition to the normal nonresponse adjustment, we will attempt to adjust the RDD and centralized CATI-only estimates for noncoverage of the non-telephone universe where necessary. The original RDD weighting does not do anything special to try to adjust for non-telephone households – the RDD c-CATI interviews are used to represent the non-telephone households. This could account for part of the differences between FHWAR and RDD shown below. The adjustment methodology for non-telephone households will be part of the research.

3.2 Preliminary Results

Preliminary tabulations show that the RDD sample yielded higher annual estimates ($\alpha = .1$) of participants and days than the FHWAR sample. Table 2 shows significant differences for key FHWAR characteristics and also for households whose members travelled.

Demographically, there did not appear to be significant differences in sex, race and Hispanic origin between the two surveys, but there was a significant difference for household income.

4. GENERAL CONCLUSIONS AND IMPLICATIONS

Some evidence has been shown that households which don't get interviewed by c-CATI are different from households which do get interviewed by c-CATI. We want our estimates to include some representation of the people who don't get interviewed by c-CATI: both people who don't have phones and people who have phones but are difficult to reach by telephone. We will continue to do research on alternative methods for adjusting for non-telephone households and c-CATI nonresponse.

5. ACKNOWLEDGMENTS AND DISCLAIMER

Thanks to all who helped with the research, reviewed the paper and helped with the presentation including Chaya Moskowitz, Dennis Schwanz, Carol Mylet, Tom Moore and Thelma Willis. Note that the authors' opinions do not necessarily reflect those of the U.S. Bureau of the Census.

Table 2
Weighted Comparisons – Key Characteristics for FHWAR and ATS

Estimate {thousands} (Standard Errors)	FHWAR	RDD	% Difference
Total Anglers	35,246 (480)	40,248 (2,200)	14.2
Days of Fishing	625,893 (19,000)	743,111 (56,000)	18.7
Total Hunters	13,975 (280)	16,642 (1,510)	19.1
Days of Hunting	256,676 (10,000)	323,253 (29,000)	25.9
Travelling Households (½ Year)	51,765 (810)	58,371 (3,200)	12.8

Table 3
Demographic Comparisons

Percent (Standard Error)	FHWAR	RDD
Male	48.0% (.3%)	47.8% (.8%)
Female	52.0% (.3%)	52.2% (.8%)
White	83.1% (.2%)	82.3% (.6%)
Black	9.3% (.2%)	9.2% (.5%)
Other Races	7.6% (.2%)	8.5% (.5%)
Hispanic Householder	7.2% (.2%)	7.0% (.3%)
Household Income Below $25,000	34.6% (.3%)	29.3% (.5%)

SESSION C-7

New Data Collection Technology: Adapting to a Changing World

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

A PREDICTIVE DIALER

C. Decoux and L. Tanguay[1]

ABSTRACT

The basic technology of the automated dialers has been created at the end of the 70's. Its main purpose was to reduce the unproductive time due to manual composition of telephone numbers that do not answer. The automated dialers could do the work of many people in less time and for a reduced cost. This technology is in constant evolution and has to meet with more demanding time and quality criteria. A new generation of these automata is the class of predictive dialers (PD). A PD has to dial telephone numbers in a predictive way, to detect "no answer","busy", "disconnected" signals, "answering machine" or others..., analysed by appropriate hardware, and to distribute the answered calls to available agents. While simultaneously, it must avoid, as much as possible, drops (answered calls for which no agent is available) and must reduce the agents waiting time (elapsed time between two interviews).

KEY WORDS: Prediction algorithms; Call progress analysis; Automatic dialer; Benefits and productivity.

1. INTRODUCTION

In the late 1970s, severe winter storms seriously damaged the power lines in areas served by the Wisconsin Public Service Corporation (WPSC). The utility's immediate problem was to call in as many employees in as short a time as possible, while saving as much time as possible for the people responsible for reaching the maintenance staff.

The system devised by the WPSC not only solved the immediate problem but also marked the initial steps in the development of automatic dialers.

Automatic dialers were originally useful in "collection" campaigns. In recent years, the associated technologies have made phenomenal advances, and dialers are starting to be used in every field imaginable (telemarketing, opinion polling, satisfaction surveys, customer service, *etc.*). What they all have in common is a need to reach people by telephone.

In this paper, we will provide an overview of automatic dialers. Specifically, we will describe the dialer we developed for use in our company's computer-assisted telephone campaign system. In the final section, we will present a model that estimates the benefits of using our automatic dialer in a telephone campaign.

2. FROM MANUAL DIALING TO AUTOMATED CALLING

2.1 Manual Dialing

What is the minimum time needed to dial a telephone number manually?

Suppose, for example, we are running a campaign with a 50% answer rate (on average, half the calls are answered). Let's consider two typical calls: for one there is no answer, while the other is completed. The minimum waiting times are as follows:

- Look up a number		2 sec.
- Dial		2-5 sec.
- Wait and listen for the call outcome		25-30 sec.
Total	(1st call)	29-37 sec.
Time between calls		2 sec.
- Look up a number		2 sec.
- Dial		2-5 sec.
- Wait and listen for the call outcome		5-30 sec.
Total		40-76 sec.

This example does not take into account the lead time the interviewer may need to look up information about the person he/she wants to reach. Hence the figures represent a best-case scenario. In addition, answer rates are usually in the vicinity of 40%. All these factors merely increase the interviewer's unproductive time. According to generally accepted estimates, the average call centre achieves an average talk time of 15 minutes per hour per interviewer in a manual environment. Hence, about 45 minutes per hour is unproductive time. A highly trained call centre might average 20-25 minutes per interviewer (Szlam and Thatcher 1996).

How can we reduce the unproductive time? The WPSC came up with the initial solution by designing the first automatic dialer.

2.2 Definition of an Automatic Dialer

An automatic dialer is a device that automatically makes outgoing calls and transfers completed calls to interviewers.

[1] Claire Decoux, Info Zéro Un, 1134 Ste-Catherine W., Suite 600, Montreal, Quebec, Canada, H3B 1H4; e-mail: claire.decoux@izusoft.com; and Louis Tanguay, R&D Director, Info Zéro Un; e-mail: louis.tanguay@izusoft.com.

There are various types of automatic dialers. Two of the most common dialing methods are preview dialing and power dialing.

A preview dialing system generally provides the interviewer with on-screen information about the potential respondent before the call is made. If the interviewer wants to proceed, he/she validates the call. The dialer then dials the number and transfers the call if it is answered. Time is saved in dialing the number and recognizing the signal on the line. Power dialing is an automated method of dialing a set of telephone numbers one after the other. It is quite similar to preview dialing.

The development of automatic dialers increased average talk time per interviewer to 30 minutes an hour (power dialing), for an average gain of 100% in interviewer productivity relative to manual dialing (Desposito 1996). Another significant advantage is an improvement in interview quality. Since interviewers no longer have to manually dial several numbers before getting an answer, they are more relaxed and better able to conduct a successful interview.

2.3 Call Progress Analysis

Automatic dialers commonly use signal recognition techniques, which first appeared in the 1970s. These techniques are used to identify the specific signals returned by busy lines, fax machines, modems, lines with no dial tone, lines that have a dial tone but are not answered, and so on. One of the first such techniques was cadence signal detection (CSD). By analysing signal cadences, it enabled the dialer to "learn" the signals used on the lines, to record them as reference patterns, to compare the recorded signals with actual signals on the lines, and to detect variations in the signals.

However, the ability to recognize the above-mentioned signal types is no longer enough. The emergence of voice mail and answering machines has made the situation a little more complicated: dialers must now be able to distinguish between human voices and recorded messages. Detection of answering machines and voice mail is usually based on signal cadence analysis. Many suppliers of such devices claim 90% accuracy (Grigonis 1995). A more realistic estimate would be an average of 70% accuracy.

2.4 Evolution: Predictive Dialers

A "simple" dialer merely dials numbers one after the other. Thus, the lower the answer rate is, the more difficult it is to reach a respondent and the greater the wait between calls will be. Waiting time is unproductive time. How can it be reduced to what is absolutely necessary? That is the challenge facing the new generation of dialers, known as predictive dialers.

How do they work? First, more than one line is assigned to each interviewer. For example, if the answer rate is 50%, two telephone numbers can be dialed at the same time for a waiting interviewer. Thus, we are "predicting" the success of outgoing calls. We can also predict interviewer availability. Specifically, one way of reducing interviewer waiting time is to start dialing even before an interviewer has finished his/her current call. Yet dialing on two lines for one interviewer is somewhat crude, and starting calls before any interviewers are waiting is somewhat risky. Moreover, reducing the waiting time between an interviewer's conversations creates a new problem: some calls will be dropped because no interviewers are available to take them. Dropped calls should be avoided wherever possible. Since it is difficult to predict with 100% reliability, they must be kept to an absolute minimum, out of respect for respondents' privacy. The generally recommended standard is to have no more than 2% dropped calls in a campaign.

As we have seen, predictive dialers are more sophisticated in their aims than automatic dialers.

Definition of a PD:

> *A PD is an automated method of making outgoing calls and transferring completed calls to interviewers. At the same time, it must minimize both interviewer wait time and the call drop rate.*

How do they work? The secret of predictive dialer technology lies in the so-called "intelligent" algorithms. Some manufacturers claim that their algorithms are based on neural network theory (Grigonis 1995). Others use sophisticated statistical techniques. All of them use mathematical algorithms that take (or should take) into account such factors as the number of available lines, the number of available interviewers, the length of an average conversation, and the time an interviewer needs between calls – in real time. All of them adjust (or should adjust) the volume of calls on the basis of those parameters. Some designers of dialers claim (Grigonis 1995) that their devices can boost interviewer talk time to 55 minutes per hour (a 400% increase in productivity over the 11 minute-per-hour average for manual dialing, and a 120% increase over the 25 minute-per-hour average for power dialing). The 55 minute-per-hour mark can be reached, but only under special conditions, such as campaign uniformity (a stable answer rate over time). More realistically, a PD can produce an average talk time of 50 minutes per interviewer.

We will now analyse the results of a typical telemarketing campaign in which the predictive dialer we developed was used.

3. ANALYSIS OF A TELEPHONE CAMPAIGN CONDUCTED WITH A SPECIAL PD

While we were developing our PD (Decoux *et al.* 1997), we simultaneously designed simulation tools to test our approaches and our results. Specifically, we devised a call outcome simulator and a PD simulator. The former simulates a telephone campaign situation while generating call outcomes and corresponding conversation times. The latter simulates the PD's activity using information provided by the call simulator and campaign specifications determined by the supervisor (cadence, number of operators, *etc.*). The simulators not only help us to analyse the dialer's behaviour in any situation but also provide a means of directly comparing versions of the same campaign conducted under different initial conditions (number of interviewers, dialer cadence, answer rate, *etc.*). Such comparisons cannot be performed in the real world since no two campaigns are exactly alike (*e.g.*, they may be conducted at different times of day). In comparing the different versions of a campaign, we are able to assess productivity variations and estimate the benefits associated with each one.

Various parameters play a role in determining how the dialer operates. Some can be set and adjusted by the user, while others can be adjusted only by the dialer (reserved parameters). The most important parameters are nA, CAD and p_{ANS}. nA is the number of interviewers assigned to the project. CAD is the cadence, a dynamic parameter ranging between 0 and 1. A cadence of 0 provides maximum safety: no calls are dropped, but interviewer wait time is at its highest. A cadence of 1 represents maximum risk: interviewer wait time is minimized, but the drop rate is inevitably very high. These two variables, nA and CAD are dynamic and can be changed by the user at any time. p_{ANS}, the project's answer rate, is a reserved parameter, which the dialer adjusts over time. The output parameters of the dialer we will study are average wait time (*WT*) and average percentage of missed connections (*DR*). The drop rate and the average wait time are computed at the beginning of the campaign. *A priori*, the output parameters depend on the parameters nA, CAD and p_{ANS}.

To illustrate how our PD works, we generated typical talk times for a telemarketing (TM) campaign. We selected TM because it allowed us to use non-zero drop rates and thus vary the cadence. Opinion poll campaigns, on the other hand, generally require a 0% drop rate, and therefore the cadence has to be zero or very close to zero. We tested two 5,000 call campaigns: one with an answer rate of 32% (TM1), and the other with an answer rate of 60% (TM2). We simulated the campaigns using the PD with 5, 15 and 30 interviewers. We considered cadences ranging between 0 (equivalent to power dialing) and 0.35. Initially, the dialer knew nothing about the campaign. It needed time to reach maximum efficiency. The length of the adjustment period depends on the number of interviewers and the answer rate: the larger the number of interviewers and the higher the answer rate, the faster the dialer responds and the shorter the adjustment period. Readings were taken after the dialer had been operating for about an hour.

The results indicated that the dialer was performing well. As the cadence rose, the number of dropped calls increased and interviewer wait time declined. The larger the number of interviewers assigned to – and active in – the project, or the higher the answer rate, the better the results for the ordered pair (*DR*,*WT*).

The relationships between the variables can be summarized as follows:

(i) for fixed P_{ANS} and nA,
$$\forall CAD_1, CAD_2 \in [0,1], CAD_2 > CAD_1 \rightarrow DR_2 \geq DR_1$$

(ii) for fixed P_{ANS} and CAD,
$$\forall nA_1, nA_2 \in N^*, nA_2 > nA_1 \rightarrow DR_2 \leq DR_1,$$

(iii) for fixed nA and CAD,
$$\forall p_{ANS_1}, p_{ANS_2} \in [0,1], p_{ANS_2} > p_{ANS_1} \rightarrow DR_2 \leq DR_1.$$

The above results are represented graphically in Figure (3.1). No matter how many interviewers there are and how long the dialer operates, it appears that the behaviour of DR and WT in relation to the cadence can be approximated by quadratic polynomial functions (coefficients of determination close to 1). Although these approximations are of no immediate use (since the polynomial coefficients depend on the number of interviewers, the project, the cadence, and so on, they represent only one particular situation and cannot be used to make general predictions), they do confirm the hypothesis that the dialer is stable and suggest that "self-regulating" methods can be used. "Self-regulating" means that the dialer no longer needs to be calibrated using the cadence. In other words, the supervisor decides the value of the desired drop rate, and the cadence adjusts itself. Even without self-regulation, the supervisor can quickly and empirically fine-tune the cadence to meet his/her needs (drop rate, interviewer wait time, number of interviewers, *etc.*).

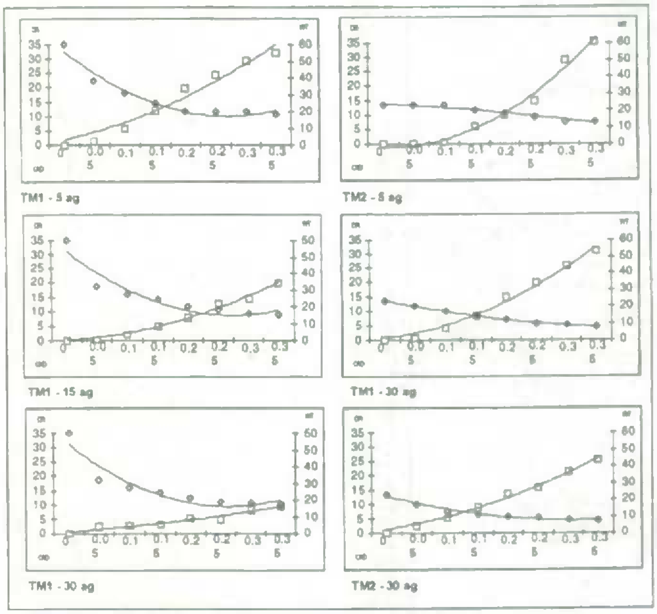


Figure 3.1 DR (squares – in %) and WT (diamonds – in seconds) as a function of the cadence (CAD) and the number of interviewers (Ag) assigned to projects TM1 and TM2.

4. CONCLUSION

Why is the predictive dialer becoming an indispensable tool for collecting data (and other things) over the telephone network?

- A predictive dialer makes the interviewer's work more rewarding.
- A predictive dialer does all the work involved in dialing telephone numbers.
- A predictive dialer can decide when to make new calls on the basis of the predicted completion time of the interviewer's current call.
- A predictive dialer increases productivity by transferring contacts immediately to an interviewer connected to a terminal.

- A predictive dialer increases interviewer talk time to over 50 minutes per hour; in a manual environment the average is between 10 and 15 minutes. In short, it collects the desired information with fewer interviewers and at lower cost.
- A predictive dialer standardizes the rules (number of rings before deciding that a call is a "no answer").

Far from being a source of stress for interviewers, the dialer allows them to relax completely between calls because it is the machine's job to find the next respondent.

To our surprise, we found during recent installations that interviewers assigned to predictive workstations did not want to go back to manual dialing.

In our next research project we will look at ways of improving the ergonomics of the predictive workstation.

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude to my colleague, François Laflamme (an expert in telephony), without whom the machine would certainly not be a dialer!

REFERENCES

Decoux, C., Laflamme, F., and Tanguay, L. (1997). Un automate d'appels cadencés prédictif. Technical report 9704, R&D Info Zéro Un, Montreal, QC, Canada. [NA].

Desposito, J. (1996). Pushing the dialing envelope. (*Computer Telephony)*, November 1996, 94-103.

Grigonis, R. (1995). Predictive dialers – Not just collections anymore. Computer Telephony (November 1995).

Szlam, A., and Thatcher, K. (1996). *Predictive Dialing Fundamentals*. Flatiron, New York.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

CALL MANAGEMENT AND THE USE OF CATI SOFTWARE BY PRIVATE SURVEY FIRMS

C. Durand[1] and S. Vachon

ABSTRACT

The paper provides an overview of computer-assisted telephone interview (CATI) software packages and how they are being used currently in private survey firms. The research draws on three sources of data: a survey of the survey firms, on-site observation of currently operations, and documentation supplied by the software manufacturers. The results show that improvements are needed in the packages' ability to manage the allocation of call attempts on the basis of call history, to determine the queue priority of uncontacted numbers, and to facilitate the tracking of operations with appropriate standard reports. Finally, the management of operations could be improved by facilitating the matching of interviewer characteristics and subsample characteristics and by providing supervisors with better training in survey methodology.

KEY WORDS: CATI; Survey methodology; Call management.

1. INTRODUCTION

This study deals with one specific aspect of the management of survey operations: the use of computer-assisted telephone interview (CATI) software to manage data collection. The research was deliberately confined to software packages used by private firms. Its main purpose is to answer two questions: whether such packages enable the firms to manage survey operations properly, and whether the firms use the programs to achieve optimal management of survey operations.

Research into CATI systems has focused on the advantages in questionnaire design and data quality. Most publications on those issues appeared about 10 years ago (Berry and O'Rourke 1988; Carpenter 1988; Connett 1990; Saris 1991; Weeks 1988). It appears that by 1988, almost all CATI programs had the minimum basic functionality, and that the situation can only have improved since then. It therefore seems pointless to go over the same ground again by re-examining functionality. The aim of this study is to take a look at the programs 10 years later, from a more qualitative point of view: how they are rated, how they are actually being used, and how they could be improved.

1.1 Background

Berry and O'Rourke (1988) and Weeks (1988) pointed out that call management in telephone surveys is not very systematic and is often based on folklore and intuition. Some issues remain unresolved (Weeks 1988): the best time for the first call, the best way of classifying appointments, call priority systems, the ability to project the amount of time required to complete a survey, and reports that can be used to make a reliable estimate of the current and expected response rate.

A number of questions requiring methodological and management decisions arise in the course of computer-assisted telephone collection operations. They can be divided into two main categories: automatic call-processing functions (call priority, calling protocols, ability to match sample characteristics and staff characteristics), and supervisory functions (control and tracking of operations).

1.2 Methodology

Information for this study was obtained from various sources. First, data were collected from survey firms across Canada (see Durand *et al.* 1997) through on-site interviews (Quebec) and mail-out questionnaires (rest of Canada). In all, 29 of the 38 companies contacted had computerized telephone survey operations, and 28 survey managers answered all the software evaluation questions and provided the name of the software package. Second, the manuals and other documentation supplied by the software manufacturers were studied. Third, an on-site inspection was carried out to determine how the supervisor(s) used each software package when it was deployed. At least one on-site evaluation was carried out for each software package. The following packages were evaluated: Interviewer, Dash, Quantime, Sawtooth Ci3 and Pulse Train.

1.3 Findings[2]

Most packages received positive ratings for *flexibility*; in one case the reviews were mixed. Ratings for *user-friendliness* were less positive, and opinion on three programs was mixed. The most frequently mentioned strengths were reliability and ease of programming, while the most common weaknesses were slowness, difficulty in programming and difficulty in changing assignments (especially for mainframe-resident software). It is important

[1] Claire Durand, Dept. de Sociologie, Université de Montréal, P.O. Box 6128, Downtown Station, Montréal, Québec, Canada, H3C 3J7; e-mail: durandc@ere.umontreal.ca.

[2] Detailed tables of the results are available from C. Durand.

to note, however, that the managers are usually not users of the operations management programs. Some functions are performed by programmers (programming of reassignment rules and questionnaires), while others are performed exclusively by supervisors.

2. AUTOMATIC CALL-PROCESSING FUNCTIONS

Managers were asked to rate four specific functions, and their responses varied. The ratings are compared with the programs' actual capabilities. *Assignment of appointments* to interviewers was considered fairly easy with PC-based software, but much more difficult with mainframe-resident programs; the managers' perceptions are more or less consistent with reality. Ratings concerning the *automatic assignment of contact schedules* were generally positive, and the programs have the functionality if the latter is assumed to mean only the general capability of assigning schedules on the basis of minimal rules (time between attempts). *Reassignment to specific interviewers* did not appear to cause any problems, whereas the *assignment of telephone exchanges* with specific characteristics seemed to be difficult for many managers. For most software packages, the latter function may in fact be hard to perform without precise information about the characteristics of the exchanges.

2.1 Priorities and Queues

Should uncontacted numbers take precedence over previous contacts? If priority is given to the previously reached (PR) sample, part of the sample may not be contacted for the first time until late in the survey operation and will thus have less chance of being surveyed. This approach assumes that initial mixing of the samples will eliminate any possibility of systematic bias. Where priority is given to uncontacted numbers, the sample is generally closed (preset number of telephone numbers), and the aim is to ensure that every number has an equal chance of being reached.

Three software packages have the capability to give priority to uncontacted numbers. In two of the three, the supervisor must intervene, either in the queue (by changing the priority assigned to the results banks) or by giving different sets of instructions to interviewers. Two packages automatically give priority to previously reached numbers, and that feature is difficult to change. This practice is apparently related to the use of quota sampling, in which the initial sample is virtually unlimited and the portion that remains uncontacted at the end of survey operations is not considered part of the initial sample.

2.2 Calling Protocols

The sample must be called at different times of day and on different days of the week in order to detect business and secondary residence numbers and to have the greatest possible chance of reaching people who are not often at home or work evenings. Weeks (1988) lists four types of call management: the *every shift approach*, in which every number is called during every shift; the *scatter approach*, in which calling times vary by day and time of day; the *contact probabilities approach*, in which the time of the next call is dictated by the probability of success; and the *priority score approach*, in which priority in the queue is based on many factors, including the number of previous attempts, the time those attempts were made, and the probability of contact. To what extent is it possible to distribute calls automatically by computer?

With the software packages studied, it is not easy to program and check a method of call-attempt distribution that takes call history into account. It is, however, easy to program the every-shift approach – if it is not already set by default – and to determine what a shift is. In most packages, it is also possible to program cumulative rules that will vary the time between contact attempts depending on the number of attempts. It appears to be very difficult, if not impossible, to determine whether all numbers have been tried at specific times of the day (daytime, evening) and of the week (weekday, weekend). In most packages, the scatter approach can only be implemented through some complex programming.[3] Some firms circumvent these difficulties by programming half-shifts or treating uncontacted numbers as appointments. More sophisticated approaches based on contact probabilities or priority scores are unavailable.

2.3 Capability of Matching Sample and Staff Characteristics

Various factors must be considered in fine-tuning the sample. While telephone numbers are usually allocated randomly, optimal management of human resources would dictate that certain exchanges or telephone books should be distributed at the beginning of the collection process or to certain groups of interviewers. Do the software packages allow easy tweaking of all operational parameters?

The packages permit some fine-tuning at the subsample level, but it is fairly difficult. The associated functions are often so complicated that users prefer to ignore them – by requiring that all interviewers be bilingual, for example – or resort to roundabout methods because they are unaware of the program's capabilities. In three packages, instructions can only be in one language; as a result, interviewers must cope with having the question in one language and the associated instructions in another.

For population surveys, a module could be introduced to inform interviewers of the location they are calling and the location's linguistic profile; to do so, system variables would have to be imported and linked to the sample file. This capability could then be used to match interviewer characteristics with subsample characteristics. In most software packages, interviewers can be grouped by specific characteristics and given assignments based on those characteristics. However, those functions are rarely if ever used, probably because of the turnover of interviewing staff or the difficulty of reprogramming the composition of the groups.

[3] One package had been programmed to count the number of attempts by type of period (daytime, evening, weekend) and to apply rules concerning maximum number of attempts; as a result, there was at least some variation in the times at which attempts were made. Another package had been programmed to get the interviewer to select the time of the next attempt from a menu of suggested periods.

3. SUPERVISION, CONTROL AND TRACKING OF OPERATIONS

3.1 Control of Operations

One of the specific purposes of telephone interview management software is to eliminate supervisor intervention in the allocation of telephone numbers. The perfect system is completely automatic, leaving supervisors free to concentrate on listening to and validating interviews. In practice, how necessary or frequent is supervisory intervention? With all the information needed to make decisions, how easily can supervisors perform the required operations?

Supervisor intervention in the queue – *i.e.*, "manual" assignment of parts of the sample that, under the rules, are not supposed to be assigned at a pre-determined time – seems to be the rule rather than the exception. Supervisors appear to have a good to very good grasp of that aspect of their work, no matter how easy or difficult it is to perform the required operations with the software package they are using. This observation supports the hypothesis that it is common practice for supervisors to intervene in the assignment of telephone numbers.

The supervisors who were surveyed were able to describe most if not all of the supervisory functions of the package they were using, and to say what purposes they served. They seemed very comfortable with the software, whatever it was. Supervisors and staff who were using particularly flexible packages displayed amazing ingenuity at "forcing" the software to do what they wanted, in some cases by circuitous means (using special codes to identify Anglophones, using time zones to close strata containing business numbers, using special appointment codes to control call allocation).

Supervisors generally had some, albeit rather superficial, knowledge of the rules governing the assignment of calls in the queue. However, they usually avoided intervening in that part of the program, even when they could access it (*i.e.*, when it was not read-only). Only one of the supervisors surveyed had the capability to intervene and actually did so.

In view of professed desire of software manufacturers and survey managers to fully automate collection management, it is striking how regularly most supervisors intervene in the "automatic" management of operations to reactivate classified numbers, classify and reassign appointments, monitor quotas and change their weights, *etc*. Most of these operations are necessary because the calling protocol rules are not specific or flexible enough or are simply ill-suited to the kind of management that the firms want to do.

There also arises the question whether supervisors prefer to intervene in the queue in order to maintain the impression that they have some control over operations. Although almost complete automation of telephone number assignment remains a goal, it will have to provide supervisors with some, possibly even subjective, control over survey operations by improving the availability of information on their progress.

3.2 Tracking Capability

Tracking is a key aspect of managing survey operations since it enables staff to act on and react to operational developments. It is especially important since the sample's characteristics – the sample frame's validity and eligibility rates – are unknown. The tracking of survey operations should provide the information needed to measure the performance of the sampling plan, monitor staff performance and forecast personnel requirements.

In most cases, producing progress reports on survey operations – status of the sample, status of appointments, current response rate, interviewer productivity – requires either advance programming or a knowledge of various complex and not-so-complex commands. As previously noted, the software packages do not always provide easy access to that kind of information. In fact, a number of firms manually input data from the call management software or statistics kept by interviewers (on paper) to database programs such as Excel or Lotus. The very fact that firms use another program suggests that call management packages lack the proper capabilities in that area, or that supervisors or programmers receive inadequate training or information.

4. CONCLUSION

CATI software packages have certainly facilitated interviewing and either improved questionnaire design or at least expanded the design possibilities. As far as the management of collection operations is concerned, however, there are serious weaknesses in the automation of call allocation and in the capacity to track operations. This is no surprise. First, before computerization, the management of survey operations had not been systematized (Weeks 1988; Berry and O'Rourke 1988). Calling protocols were variable and based on intuition (Weeks 1988). Second, private companies frequently use quotas (Durand *et al.* 1997). In most if not all software packages, quota management is easier and more sophisticated than the management of contact attempts, which is needed for samples with callback. The issue of reports and tracking is more complex since the packages seem to provide fairly good tracking and flexibility in allowing the user to determine what information the reports should contain. The fact that some firms use other tools suggests that the tools provided are inadequate, that they are considered too complicated to use, or that training is deficient. None of the packages seems to be able to help in planning the resource requirements for collection operations. In addition, the kind of fine-tuning required to match interviewer and sample characteristics is difficult to carry out. The packages have the capability to improve this aspect of call management, but the functions are seldom used. This may be due in part to high turnover among interviewing staff, frequent schedule changes, and the complexity of this kind of matching.

How can the software packages be improved? Improvements in two areas are essential. First, *introduce modules that will facilitate the use of the scatter approach*. The functions used to *match interviewer and sample characteristics* should also be improved. Second, *introduce standard reports* that would automatically provide estimates of the expected response rate or the cooperation rate, the estimated number of call-backs required at selected times, and productivity data that could easily be transferred to current database packages.

The supervisors who handled the day-to-day operations of telephone surveys seemed to have a good understanding of what they had to do and how to go about it. However, it is not clear whether supervisors have the theoretical training they would need to understand why the operations are carried out one way and not another. The gap between professionals, methodologists on one hand and technicians on the other, which has persisted and perhaps even widened since the advent of CATI software, is probably an obstacle to the improvement of call management. Staff generally receive on-the-job training, and each firm produces its own reference manuals for supervisors and interviewers. Training appears to be piecemeal and fragmented, as each staff member knows only the parts of the system that he/she uses regularly. Surely it is time for the firms to acquire the tools to provide survey personnel with professional training. Both the firms themselves and the survey industry as a whole would certainly benefit.

Finally, it is worth repeating that continuing research is needed to improve and systematize call management practices and thereby maximize response rates.

ACKNOWLEDGEMENTS

We are grateful to the Social Sciences and Humanities Research Council of Canada (SSHRC), which provided a grant to fund our research. We would also like to thank the survey firms, especially for their cooperation during the observation phase.

REFERENCES

Berry, S.H., and O'Rourke, D. (1988). Administrative designs for centralized telephone survey centres: implications of the transition to CATI. (Eds) R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, and J. Waksberg, *Telephone Survey Methodology*, New York: John Wiley, 457-474.

Carpenter, E.H. (1988). Software tools for data collection: microcomputer-assisted interviewing, *Social Science Computer Review*, 6, (3), 353-368.

Connett, W.E., Blackgurn, Z., Gebler, N., Greenwell, M., Hansen, S.E., and Price, P. (1990). A report on the evaluation of three CATI systems. Survey Research Center, Institute for Social Research, University of Michigan, 99.

Durand, C., Tanguay, I., and Vachon, S. (1997). La gestion de la méthodologie dans les firmes de sondage au Québec et au Canada. *65 ème Congrès de l'ACFAS*, Université du Québec à Trois-Rivières, May 16, 1997.

Nicholls II, W.L. (1988). Computer-assisted telephone interviewing (Eds) R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, and J. Waksberg, *Telephone Survey Methodology*, New York: John Wiley, 377-385.

Saris, W.E. (1991). *Computer-assisted interviewing*. Newbury Park: Sage.

Weeks, M.F. (1988). Call scheduling with CATI: current capabilities and methods (Eds) R.M. Groves, P.P. Biemer, L. E. Lyberg, J. T. Massey, W.L. Nicholls II, and J. Waksberg, *Telephone Survey Methodology*, New York: John Wiley, 403-420.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

NETWORKED HAND-HELD CASIC

S.M. Nusser and D.M. Thompson[1]

ABSTRACT

Computer-assisted survey information collection (CASIC) has grown to include hand-held pen-based architectures for collecting data. These architectures are most efficient if they are connected to a centralized data base server that functions as a repository for sample data. Recent research in data collection methods for a national natural resource survey has led to a system that utilizes Newton MessagePads clients with software to collect data and transfer it to and from a centralized server. Sample data can be downloaded or uploaded, and the system can be used to facilitate CASIC software updates. In addition, the centralized server concept has been broadened by using the Web to monitor survey progress. This system is currently being used to collect data on hundreds of variables for 300,000 PSUs and 800,000 sample points. We discuss designs for CASIC systems and computer-assisted survey instruments, and outline current and future extensions including GPS and GIS functions and access to the Web via the hand-held computer.

KEY WORDS: Personal digital assistants; Pen computing; Mobile computing; Computer-assisted.

1. INTRODUCTION

The U.S. Department of Agriculture and the Iowa State University Statistical Laboratory have collaborated for a number of years on a large natural resource survey in which data gatherers require a great deal of mobility. In past surveys, computer-assisted data collection has involved desktop computers, which were used to enter data into electronic data collection forms. Recent advances in hand-held computing have led us to investigate alternative methods of accomplishing data collection with a mobile workforce. The purpose of this paper is to describe our experiences with developing large and small scale applications of hand-held CASIC.

We begin by describing survey settings in which a mobile data gathering system is useful, and discuss current hand-held computing features that are viable in these settings. We then outline what our experiences with developing hand-held CASIC systems have revealed, and areas we expect to explore in future research.

2. NATURAL RESOURCE SURVEYS

Our original objective was to develop computer-assisted survey information collection methods for a large survey called the National Resources Inventory (NRI). The NRI is a national survey program of the U.S. Department of Agriculture, designed to monitor conditions and trends for natural resources. Its primary purpose is to support agricultural policy development, and it is used in the construction of farm bills in the U.S. The NRI also supports agroenvironmental research objectives, such as modeling economic and environmental effects of alternative agricultural policies or bio-physical processes such as pesticide leaching. Special topic investigations are conducted regularly as part of the NRI program, for example, to monitor the effects of recent farm legislation on conservation practices.

The NRI is a stratified two-stage area sample of U.S. lands, consisting of about 300,000 primary sampling units, which are usually 160 acre (64 hectare) square area segments, and about 800,000 secondary sampling units, or points. The survey is longitudinal, with data collection occurring every five years on nonfederal land.

Data collection is based primarily on photo interpretation, with effort devoted to abstraction of office records to obtain information from conservation plans and soil surveys. Field visits are made to sample units when data are not available through standard materials, or when it is required for special studies. At the primary sampling unit level, areas of polygons defining specific land uses are recorded, as well as the length of linear features such as streams. Within the primary sampling units, usually three points are selected. The variables collected at each point include land use classifications, agricultural practices, soil characteristics, and other natural resource attributes, such as habitat and wetland information.

3. MOBILE SURVEY SETTINGS

The survey setting for the NRI requires the data gatherer to be mobile at a number of different scales. Within an office, data gatherers need to be able to move between photo-interpretation stations, paper files, and computers that support various functions, such as access to GIS tools and to the Web. On a larger scale of mobility, data gatherers

[1] Sarah M. Nusser, Department of Statistics, 220 Snedecor Hall, Iowa State University, Ames IA 50011-1210 U.S.A.; e-mail: nusser@iastate.edu; and Dean M. Thompson, Natural Resources Inventory Analysis Institute, USDA Natural Resource Conservation Service; e-mail: deano@iastate.edu.

may need to go to offices of other agencies to abstract information from records, or they may need to go out to the field in order to observe conditions. An additional feature of the survey setting is that multiple data gatherers may work on the same sample unit. These data gatherers may not be at the same site, and thus, a method of sharing the sample units and their associated data is required.

Other features of the survey create challenges in developing forms with hand-held computers. Historical data are very important in locating samples and in classifying current conditions and parameters. Because of this, historical data need to be available within the context of the computer-assisted survey instrument. Numerous edit rules are required to check the consistency and accuracy of the data. Finally, there is increasing pressure to rapidly develop surveys in response to special issues as they arise. A computer-assisted survey system would ideally include features that allow survey instruments to be created quickly and that would allow key question modules to be re-used in new survey instruments.

While data collection for natural resource surveys can be quite complicated, numerous challenges exist in other settings as well. For example, when conducting in-person interviews, particularly with mobile respondents such as farmers, hand-held computers could be quite advantageous. Hand-held computers can be useful in enumerating primary sampling units. Of particular interest is capturing household information in a form that retains spatial relationships among housing units using GIS capabilities on hand-held computers. Hand-held CASIC can also be beneficial in a number of data collection settings where information is being abstracted from non-human sources, such as in pricing surveys and office record abstraction. The main feature of interest is site-to-site mobility.

4. HAND-HELD COMPUTING

The field of hand-held computing has greatly expanded over the last year or two. It is now possible to purchase hand-held computers that are inexpensive, lightweight, and have excellent graphical user interfaces. Several models are pen-based. A few support handwriting recognition. Many hand-helds are equipped with ample storage capacity, as well as diverse communication options. In the past couple of years, commercial off-the-shelf (COTS) forms development software, as well as custom forms development services, have become available for hand-held computers.

Our principal hand-held CASIC experience has been with the Newton MessagePad running the Newton Operating System (NOS) 2.0. The Newton has all of the features mentioned previously, and its operating system is object-oriented. The object-oriented operating system is a great advantage in the survey setting because a number of standard forms development features are already present in the operating system.

5. HAND-HELD CASIC SYSTEMS

Newton technology provides an environment in which it is possible to develop functionality that is commonly seen in traditional PC-based CASIC systems. It is possible to develop a computer-assisted survey instrument (CASI) with flexible question and answer formats. Furthermore, hand-held CASI's clients communicate with a central data base server.

The hand-held architecture of Newton technology offers enhancements over traditional PC-based CASIC systems. Software is available that allow GPS receivers to interface directly with the survey instrument on the Newton. For natural resource surveys, GPS connectivity is important in locating samples and recording sample locations. GIS tools are also being developed. GIS presents some intriguing opportunities in many kinds of survey settings, including segment enumeration as noted earlier. In biological surveys, the GIS interface can be used to record plot features and implement random sampling within the plot. For example, streams within a plot can be recorded and a point randomly selected for observation. GIS objects on the hand-held can be transferred to a GIS interface such as ArcView.

6. SYSTEM DESIGN

In its most sophisticated form, a hand-held CASIC system has components similar to those of a PC-based system. It is based on remote computers that connect to a central data base service. Our approach to developing hand-held CASIC systems has been to include materials that are required by the data gatherer on the hand-held computer, while making alternative views of the survey data and summaries of progress available to the survey manager via a centralized service.

The central data service stores and serves all survey data, and also supports messaging and CASI software updates. In addition, it serves reports to monitor survey progress, typically to a survey manager's PC system. In the systems we've been developing recently, Web-based training and a Web-view of the data have been incorporated to support remote training, monitoring, and post-data collection editing from a Web browser. While the Web-based materials are currently available via a browser, we are also interested in working with the Web environment on the hand-held computer. We anticipate using new tools, such as Java, as a means to develop a single product for many architectures.

For the systems we've been designing, it has been advantageous to provide a number of methods for connecting the hand-held unit to the central data service. In field situations, a cellular phone can be used to reach the central service. In the office, a wireline is used to connect directly to the central service, or to the central service via an Internet service provider or corporate intranet.

7. CASI DESIGN FOR HAND-HELD COMPUTERS

Because hand-held computers have limited screen size, emphasis has been placed on developing simple user interfaces for CASIs and supporting materials. Standard tools are readily available on NOS, such as pick-lists, check boxes, and radio buttons. We have found it possible to develop complex questionnaire formats that are easily interpreted with smaller screens. For example, we've been able to create complicated displays to address special needs, such as displaying historical data for related variables, using a variety of hybrid solutions that involve horizontal and vertical scrolling grids.

In addition, simple interfaces for collecting data on multiple reporting units per sampling unit can be developed, as is required for household rosters or for recording attributes of water bodies in a sample area. For this purpose, we have used paged sections, in which each page contains the variables to be entered for the reporting units within the sample unit. Grids and pages are examples of electronic formats that mimic paper paradigms that are familiar to data gatherers.

The emphasis on a simpler user interface has led to extensive nonlinear navigation capabilities on hand-helds. It is relatively simple to develop hypertext links to specific sections from a readily accessible overview of the CASI, to establish buttons for commonly needed links to various sections in the survey instrument, or to link parts of the survey instructions on the hand-held to the relevant section on the CASI.

In sophisticated applications, it is also possible to include extensive edit rules to check for legal entries for each variable and consistency among variables, or to invoke calculations. Convenient error messaging systems have been developed to display the results of such checks. The ability to place edit rules and checking at the fingertips of the mobile data gatherers during the data collection phase has reduced our post-data collection editing volume by roughly 75%.

Many kinds of on-line support can also be included on hand-held computers, such as survey protocols, definitions, more detailed explanations of error messages, and maps.

8. CASI DEVELOPMENT

Our experience with the creation of CASIs has ranged from COTS to custom developed software. Because of the complicated nature of our CASIC requirements, we began by working with customized COTS software. This allowed us to develop a very sophisticated CASI, using TCP/IP-based communication protocols with the central service. However, it took a fair amount of time to work out the interface and communications.

We were eventually faced with fielding a couple of surveys where we did not have the kind of time required to develop a full-featured CASIC system. For these surveys, we relied on a fully COTS system. This approach gave us the time advantage, but we were forced to make compromises in the complexity of the system which led to increased post-data collection processing. Very few edit checks were incorporated, and only the simplest questionnaire formats could be used. Interaction with the central service consisted of sending PCMCIA storage cards back and forth or returning data via e-mail. Nevertheless, because most responses were pre-coded, it was a net gain to have electronic data collection forms.

Recently, we have started developing computer-assisted survey instruments using the developer's package, Newton Tool Kit. Newton's object-oriented operating system made this fairly efficient, and we were able to regain some of the sophistication found in the customized COTS solution. At the same time, it took a lot less time to construct the survey instrument than when we used the customized COTS solution because we were in control of the design process.

During these experiences, we developed GPS/hand-held computer intercommunication and an e-mail interface for returning data in simpler situations. In addition, we created systems to support concurrent surveys that required access to data associated with a common set of sampling units.

Ultimately, the choice of CASIC system will depend on the survey requirements and the resources available for development. We see two extremes in this choice. For large complex surveys with lengthy planning horizons, as is the case with the NRI, it is possible to develop an excellent customized CASI with extensive edit rules and sophisticated question and response formats. Key modules can be developed so they can be re-used in future surveys. Modern client and server technologies can be used to knit together the mobile CASI's and central data base service. For this kind of a system, complexity and development time can be quite high, and a multi-disciplinary staff is required. But there can be large pay-offs in data quality and convenience in handling large volumes of data.

Alternatively, for simpler surveys, it's possible to use COTS forms development software. However, a price is paid with limited calculations, edit rules, and response formats. Essentially, a single-use CASI is developed. The survey must also rely on fairly simple remote or local information exchange. System complexity, development time, and requirements for technical staff are much lower. This is a cheaper solution, appropriate for situations in which non-programmers are the developers of the CASI. However, data processing time is higher under this setting.

Our initial customized COTS approach is an intermediate option. Currently, the hand-held CASIC market appears to be too small to generate much interest on the part of COTS forms vendors. While our interaction with a COTS vendor initially benefited both our survey program and their more generic forms development software, ultimately the larger markets, such as medical applications, made our work less attractive to the company. We hope this situation improves as larger survey organizations embrace hand-held technologies.

9. CONCLUSIONS AND FUTURE WORK

In summary, it is now possible to develop fully functional hand-held CASIC systems for large and small surveys. Those systems may be sophisticated client server solutions with a complex survey instrument, or they may be fairly simple for smaller surveys.

Traditional survey materials can be incorporated into the system in new ways. Resource materials can be loaded onto the hand-held, such as providing on-line instructions with hypertext links. We can also consider the possibility of providing GIS-based maps on the hand-held along with GPS links to locate sample units. Expanded electronic capabilities now exist in remote settings. By coupling a cellular phone with the hand-held, it is possible to exchange information and to use e-mail, fax, and Web-based tools from virtually any location. The Web also presents some interesting opportunities in providing centralized training, monitoring, and editing functions in remote locations.

We have a number of areas that we are continuing to pursue. First, we are interested in developing alternative methods of data exchange in simpler settings using NOS. A fairly simple solution relying upon PC connectivity will be required for smaller surveys. Second, we are currently exploring the use of GIS tools within the computer-assisted survey instrument to record and process information, as well as to create objects that can be passed to another system. And finally, we expect to pursue the use of the Web in the hand-held environment as an alternative method of providing supporting materials and survey management information to the data gatherer.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

ELECTRONIC DATA REPORTING

M. Ménard and G. Parent[1]

ABSTRACT

The Operations Research and Development Division (ORDD) has been exploring the challenges and opportunities that modern communications technologies present for the electronic collection, capture and integration of survey data. In partnership with the Operations and Integration, International Trade and Labour divisions, the ORDD has conducted pilot projects to gather data electronically and eliminate data entry. Two different collection methods were used for these pilot projects: use of an electronic questionnaire in the International Trade Division project, and preparation of an electronic file in the Labour Division project. The businesses participating in the surveys had a choice of ways to submit their data: electronic data interchange (EDI), file transfer, or electronic mail over the Internet. The projects required the development of a technical communications infrastructure and the use of encryption techniques. In our paper, we will share the lessons we have learned in this area, spell out the options available under the current infrastructure, and suggest avenues for future development.

KEY WORDS: Electronic data reporting; Infrastructure; Security.

1. INTRODUCTION

The recent years have witnessed the wide scaled adoption of personal computers in the workplace. The increasing user friendliness of operating environments and software suites has broken down cultural resistance by facilitating their exploitation. The new medium has multiplied the means of communication available to the general population by giving rise to electronic mail and the World Wide Web through the Internet. These developments have expanded our opportunities to reduce response burden in our data collection activities.

The advent of electronic data reporting is a continuation of our periodic introduction of new collection tools. In effect from the paper based questionnaire we have evolved to Computer-assisted Telephone Interview (CATI), Computer Assisted Personal Interview (CAPI), Imaging, Intelligent Character Recognition and now to Electronic Data Reporting. Our objective has not been to replace one by the other but rather to develop the best set of collection tools for a given survey.

2. MAJOR BENEFITS

Electronic reporting is being pursued in light of the benefits that can be gained. These can be expected both for the users and Statistics Canada.

2.1 Users

In many instances the drive for this method is coming from the respondents who wish to avail themselves of a reporting option which meets their technical capabilities and today's communication technologies. It is also a development, which is consistent with Treasury Board's statement that electronic commerce will be the preferred way of doing business for the Canadian Government in the future. Of even greater significance is the potential for reduction in response burden with the eventual possibility of importing data from corporate databases to electronic questionnaires transmitted from and to Statistics Canada by electronic means.

2.2 Statistics Canada

Cost efficiencies in collection will arise as we eliminate the physical intervention for delivery and retrieval of the paper questionnaire. The data capture cost will be lowered as it is passed on to the respondent whose incentive will be the reduction in response burden. The cost of editing will also go down as we gradually incorporate edits into the electronic questionnaire, which can lead to further reduction in follow-up.

We can expect an increase in *data accuracy* with the inclusion of edits to which the respondent can react to in real time. The crystallization of the principle of "Get it right the first time" rendered possible by this method, amongst others, will improve the quality of the source data.

As it is quicker to deliver questionnaires and retrieve them through electronic means then it is by Canada Post or Courier, timeliness will be positively impacted. The potential reduction in the volume of follow-ups will allow for earlier analysis and therefore dissemination. This will also be of benefit to the user.

3. SECURITY

3.1 Introduction

The major challenge confronted in establishing electronic data reporting is the satisfaction of the stringent

[1] Mario Ménard and Guy Parent, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

security requirements mandatory in the collection of confidential survey data while keeping cost to a minimum. This is currently achieved by applying the policy ,in effect since 1990, which outlines the implementation measures regarding the use of computers and facsimile equipment for receipt of survey data from respondents. It does not describe how it should be done but rather provides guidelines to be followed.

The policy declares that the Department recognizes the possibilities of unauthorized access during the communications process and requires adherence to administrative and technical procedures that are intended to minimize the risks involved. It involves encrypting data at source and throughout the transmission process with decryption taking place in Statistics Canada's secure network. If encryption is not implemented at source the respondent must sign a "Letter of understanding" which states that there is a risk of unauthorized access to the information during the electronic transmission of data and that the risk is acknowledged and accepted.

The policy is expected to evolve to reflect technological advancements in browser technologies particularly in the area of security.

3.2 Secure Transaction

The basic components of a secure transaction are the following:

- Privacy means that no one can read the file being sent, except for the sender and the intended recipient.
- Authentication validates the originator of the message. In effect, it verifies that the sender is who he says he is.
- Finally, integrity means that the file cannot be altered during transmit.
- Non-repudiation means the sender cannot deny creating the messages.

According to industry specialists, the only approach that guarantees all theses elements is public-key technology, also called public-key cryptography. On the other hand, key management is the most difficult aspect of adding security to an application and has been a roadblock to widespread use of encryption technology.

The contracting out of portions of the communication infrastructure to the Public Works Government Services and their Government Telecommunications and Informatics Service (GTIS) to make use of their existing electronic expertise and infrastructure has allowed us to minimize costs while adding additional security layers. Respondents send their data to GTIS before it gets to STC for most options. A secure environment for the data transmitted between GTIS and STC has been established and is currently available using Entrust software products.

Entrust is a family of software products for encryption and digital signature on client/server networks with fully automated key management. At this point in time, it is too early to determine the time required and the feasibility of implementing this infrastructure in our EDR facilities. In the short term, it is recommended that the current policy be applied to any electronic transmission and that work continues with EDP security to put in place a complete local infrastructure.

4. CURRENT EDR ENVIRONMENT

4.1 Evolution

The evolution to a full electronic reporting capability at Statistics Canada has already witnessed some significant achievements.

EDR Centre in O&ID – The set up of an EDR Centre in Operations and Integration Division (O&ID) was an important milestone. The Centre currently collects data from about twenty-five companies from various client divisions, using a modem to modem connection. Pre-arrangements for transmission are made with the institutions. The file transfer is monitored on a network B work station and once complete is copied to diskette and deleted from the hard disk. The diskette is then moved to a network A Workstation and loaded. The respondents are notified by telephone of a successful or unsuccessful transfer.

Diskette based Questionnaire – The implementation of a Computer Self-Administered Questionnaire (CSAQ) approach in the Annual Retail Trade Survey, P13 and Steel Survey was another major step. The approach was realized through the development of a diskette FOXPRO based system known as the Personalized Electronic Reporting Questionnaire System (PERQS). The application incorporates edits including historical data and the diskette base questionnaire is therefore encrypted when sent to the respondent and when returned to STC. Security is further enhanced by the questionnaire being delivered and returned via courier. An interesting feature of PERQS resides in the respondent's possibility of importing data from an Excel or Lotus spreadsheet.

Electronic Files and Forms – In this fiscal year a communication infrastructure was developed and implemented by Operations Research and Development Division to allow, for the first time, an end to end solution to electronic reporting for the Business Payroll Surveys (BPS) of Labour Division and the Exporter Declaration Form (B13A) for exporters with the International Trade Division. The BPS uses an ASCII file while the B13A is an electronic form (The form is also available on diskette). The mode of transmission offered ranges from the File Transfer Protocol (FTP) via the Internet, Electronic Data Interchange (EDI) and E-Mail attachment. A structured data transfer (SDT) which makes use of the secure connectivity offered by a Value Added Network with EDI while removing the need to adhere to costly EDI standards and protocols, was also developed for the BPS.

Data transfer modes – FTP – Respondents using this method must be Internet enabled and have file uploading software installed available as a download or by diskettes from the Canadian Automated Export Declaration site of ITD. The B13A is encrypted at source using PGP (Pretty Good Privacy). The files are uploaded via the Internet to the

GTIS web server at Public Works. As files arrive at GTIS they are moved to a server on the secure side (behind the firewall) of the Public Works network from where they are retrieved by STC.

EDI & SDT – The Labour Division provides its' respondents with a diskette containing the connectivity software to the Value Added Network. The required data set is transferred over dedicated lines to a Value Added Network. It is then retransfered, again over dedicated lines, to the Public Works mainframe or Government Electronic Data Interchange (GEDIS). The data is then moved to the same STC pick up server referred to under FTP.

E-Mail – Respondents availing themselves of this approach send the data file as a mail attachment to a Statistics Canada mail box where it is currently redirected by ORDD to the destinator mail box. It is not the preferred option due to problems with file size, reliability, and differing mail standards.

Retrieval by STC – The need for retrieval from GTIS to STC applies to data collected by FTP and EDI or SDT. On a set frequency which varies by survey application, Operations Research and Development Division opens a unidirectional channel from the public side of the STC network to the pick up server site on the secure side of the Public Works firewall. The data files are then pulled in to the STC server on network B. While in transit between Public Works and STC the data is encrypted with Entrust, supported by the Government of Canada, and will remain encrypted until it is on the secure A network. The data on the B network is then sent to a server A network using Kyber Win software to penetrate the firewall. The data is again moved from the initial network A server to a second network A server; this is done so that data decryption can only occur in an environment that has no physical connection to the public network *ORDD Processing* – The core processing constitutes in the switching over from the Public Works secure to the STC public network and from there to the STC secure network. Once the data is in a secure environment it is virus checked, verified for completeness, verified for duplication and then decrypted. A message is then sent to erase all validated files sited on the various servers through which they have been routed while a request is sent to retransfer all damaged files. The validated files are then pushed to the appropriate subject matter divisions. Steps are currently being undertaken to automate the reminder to non-respondents process.

A toll free help line is available to respondents encountering difficulties in transmitting their data. The help centre analyzes the nature of the problem and routes a query for resolution to the appropriate party.

5. ISSUES

As we move to a large scale electronic reporting capability at Statistics Canada several issues will need to be resolved in order that we continue to progress.

We will have to develop and institute *standards* for the design and appearance of electronic forms, questionnaires and collection web sites. These standards should control the proliferation of methods and approaches which if unchecked can prove costly; ensure consistency in the image that Statistics Canada projects to the public; ensure complementarity development. It is also considered important that standards be not restrictive to the point of impeding creativity and innovation.

The on line completion of questionnaires over the Internet, with historical edits, will require the establishment of a *public key management* infrastructure with authentication and end to end encryption using an approved Government of Canada standard, in this instance ENTRUST. We need to determine the merits of contracting out to a specialized government agency the infrastructure management versus developing and managing an in house infrastructure. The whole question of licensing of encryption keys to respondents and associated costs requires rationalization.

We must deepen our *relationship with software developers* and vendors. This relationship will aim to make available to our respondent's packages that can facilitate the mapping from databases of desirable data elements for automatic integration to electronic forms and their further integration to Statistic Canada databases. We must put our requirements and specifications in the private sector without creating any false perception of the development of a strategic alliance with any particular software developer.

The processes associated to electronic data reporting are new. The impact on *existing survey processes* will have to be carefully measured and integration sensitively carried out.

Constant monitoring of accelerated *technological changes* being incorporated into our respondent's environment will be required to continue to satisfy their preferred way of doing business.

The hours of operation and the precise definition of service for the assistance centres or help lines, maintained for respondents availing themselves of EDR, will have to be established.

6. FUTURE DEVELOPMENTS

6.1 Start Ups

Operations Research and Development Division has recently established new relationships with the Centre for Justice Statistics , the Transportation Division, and Revenue Canada Taxation to develop EDR applications. These applications will channel their data through the infrastructure established between the Government Telecommunications and Information Services of Public Works and Government Services, and Statistics Canada. The Division will also proceed to convert businesses reporting to the existing EDR centre in O&ID to the same communication infrastructure.

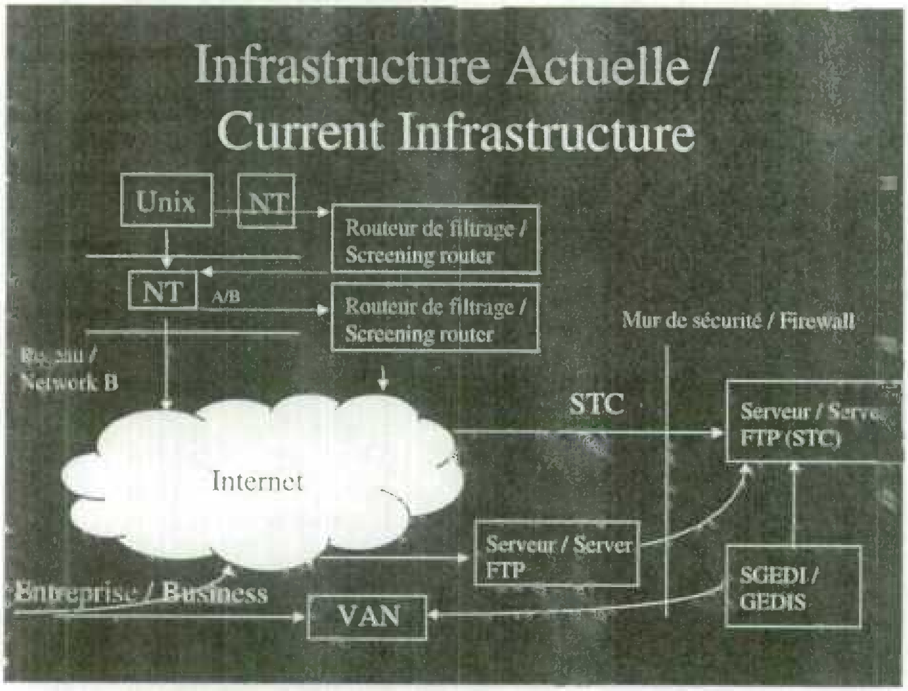
6.2 Near Future

In the near future our energies will be focussed on offering an electronic reporting option to the Unified Enterprise Statistics respondents starting with the pilot. This will consist in an electronic questionnaire download site for all of part one, two and three questionnaires in addition to a form completion assistance site. The applications will be generally based on the Canadian Automated Export Declaration (CAED) of International Trade Division model. The respondent will download connectivity and file uploading software with the electronic questionnaire. Procedures will be developed to authenticate and validate respondents in the absence of a public key infrastructure.

6.3 On the Horizon

Further down the road we anticipate the resolution of the remaining security issues associated to electronic reporting particularly as they relate to the Internet. At that point we will be ready to enter the world of Web collection were the respondent will access a web questionnaire with edits on an STC site and fill it out in real time with parallel encryption. The respondent will also be able to navigate through hyperlinks to concepts and definitions related to the relevant survey and to highlights of published previously released data to put into context his or her response.

Finally the pursuit of commercial software for our major corporate respondents to map and integrate data of statistical interest from their databases will bear fruit and become widely available.



Infrastructure Actuelle / Current Infrastructure

Unix

NT

Routeur de filtrage / Screening router

NT

A/B

Routeur de filtrage / Screening router

Réseau / Network B

Mur de sécurité / Firewall

STC

Serveur / Serve

FTP (STC)

Internet

Serveur / Server FTP

SGEDI / GEDIS

Entreprise / Business

VAN

Figure 1. Data Transfer Process Chart

SESSION C-8

Tackling Response Burden

REDUCING STATISTICAL BURDENS ON BUSINESS: DEVELOPMENTS IN THE UNITED KINGDOM

A. Machin[1]

ABSTRACT

In the United Kingdom over recent years there has been a strong drive to ease the demands on the public resulting from government requirements generally. In particular there has been downward pressure on the compliance costs of business surveys. This paper describes experience in the UK of measures to reduce statistical burdens, especially through the use of existing administrative data sources, development of business registers and electronic data collection methods. A recent independent study identified savings in compliance costs to business of government surveys of up to 26 per cent without impairing the quality of essential statistics. Future possibilities for using existing data sources and collecting data directly from businesses' own information systems are described. Developments such as special modules in commercial business software packages are particularly important given the potential for extracting statistical data by electronic means with minimal burden on the companies concerned.

KEY WORDS: United Kingdom; Statistics; Reducing; Burdens; Business.

1. INTRODUCTION

Statistics of business are vital for running the economy and business generally, yet the burden of government surveys is relatively small overall. The compliance costs to UK business amount to some £60 million per year, 0.01% of gross domestic product. The load of statistical surveys on business is also tiny in comparison with the costs of complying generally with administrative requirements of government. The burdens are nevertheless perceived by some businesses to be considerable and merit serious attention. It is important to minimise the burdens not only from the business point of view but also for the quality of the statistics. It is vital to focus on respondents to get co-operation and thus obtain reliable data.

The Survey Control Unit (SCU) of the Office for National Statistics (ONS) was established in 1968. Successive Prime Ministers have since issued instructions on the control of statistical surveys to Ministers in charge of departments to ensure rigorous assessment and review of surveys The rules have become progressively tighter, particularly in respect of surveys of business, as a result of deregulatory pressures. The SCU is responsible for ensuring that the control procedures are followed and for assessing new survey proposals and reviews as required. It aims to promote necessary surveys of high quality, prevent bad or unnecessary surveys and ensure that burdens on data providers are minimised. This is a matter of balancing the burdens on data providers against the benefits. The unit monitors and reports on survey activity and the compliance costs to business. It also promotes best practice for the conduct of surveys.

There have been various initiatives in the UK to cut statistical surveys, notably during the 1980s. Some growth in survey activity returned in the early 1990s, partly to rectify some of the deficiencies created earlier and also to meet new demands. A recent major review has led to a further reduction in the burdens although there is now more emphasis on maintaining the quality of essential statistics. The measures to reduce burdens are now better focused as the monitoring of compliance costs has revealed how the burdens on business are concentrated among a few major surveys. The 25 largest regular business surveys, some 5 per cent by number, account for some 88 per cent of the overall compliance costs.

The burdens on business resulting from surveys are varied and complex and there is correspondingly a variety of possible methods to reduce them. Many of these do not threaten the quality of the statistics produced. In the UK recent savings in compliance costs have been mostly achieved through the streamlining and simplification of surveys, more efficient sampling and the development of business registers which have enabled more effective use to be made of administrative sources of data already available to central government. Minimising the burdens on data suppliers has become a key objective of the ONS and of the wider UK Government Statistical Service. Major surveys of business are now subject to compliance planning whereby compliance costs are kept within limits agreed each year with Ministers on a three-year rolling basis. There is also a drive to improve information about available government statistics, encourage wider access and greater use, and promote coherence and harmonisation of data, under the *'wider agenda'* programme. A key component is the development of a new integrated database (*StatBase*) of statistics drawn from the whole range produced by government. This should help to ensure that existing data sources are used to the maximum extent.

[1] Andrew Machin, Head of Survey Control Unit, Office for National Statistics, 1 Drummond Gate, London, SW1V 2QQ; e-mail: andrew.machin@ons.gov.uk. A longer version of this paper is available which describes the nature of statistical burdens and the variety of measures to reduce these in more detail.

2. THE OSMOTHERLY REVIEW

An independent study led by Edward Osmotherly (UK Government 1996) was conducted in 1996 to look at how the burdens of government surveys on business might be reduced, especially for small and medium-sized enterprises. This followed a recommendation in 1995 by the former Deregulation Task Force (a Government-appointed group mainly representing business interests). The recommendations are now being implemented.

The Osmotherly report identified savings in compliance costs to business of up to 26 per cent (£17.6 million at 1995 prices) which are expected to be achieved by the year 2000 without impairing the quality of essential statistics. Some initiatives are already in progress, for example measures to streamline surveys, reduce overlaps between them, sample more efficiently and make more use of electronic data collection methods. A number of recommendations are being followed up to simplify specific surveys including (subject to negotiation) Intrastat, the European Union system for measuring cross-border trade.

A considerable part of the estimated savings is to be achieved, in partnership with software houses, through the development of commercial accounting packages to extract reliable statistical data for survey purposes automatically from existing company records. Trials have been set up. A high-level team, including several representatives of business, has been formed to see what can be done to speed up progress with such initiatives and has begun to identify issues which need to be tackled.

A major recommendation of the Osmotherly report is that samples for business surveys should, wherever practical, be drawn from the Inter-Departmental Business Register (IDBR) (see Perry 1995). The IDBR holds identification details and basic statistical information such as employment, turnover and industrial classification of individual businesses and company structures drawn mainly from administrative sources, in particular from the Value Added Tax (VAT) register, the Pay As You Earn (PAYE) employees' income tax and the company registration systems. The IDBR is an important tool enabling more efficient and more co-ordinated sampling and it avoids overlaps in the collection of basic statistical information. It enables survey burdens to be shared more equitably between businesses as well as helping towards compiling more reliable and consistent economic statistics.

Use of the IDBR will be vital in working towards the introduction of guaranteed *'survey holidays'* for small firms as the Osmotherly report also recommended. The ONS has already introduced such a guarantee from 1997 for firms employing fewer than ten people, whereby if they are included in a statistical survey they are told in advance for how long they will be expected to participate. They will not subsequently need to take part in another ONS survey for at least a further three years. This guarantee is included in a Business Charter (Office for National Statistics 1996) which explains what service business can expect from the ONS. This type of approach does not reduce the overall burdens on business but it is very important in the interests of fairness and in terms of perceptions of business. Other departments have been considering their own guarantees where possible in line with the Osmotherly recommendations. An inter-departmental committee has been considering how to maximise the extent to which this approach can be co-ordinated across all government surveys and extended to cover firms employing up to 25 people as far as practicable. Such developments present a number of challenges for government departments.

Another key development recommended by the Osmotherly report, is that annual compliance plans (as recently introduced in ONS) (Office for National Statistics 1997) are now being produced by all departments with major responsibility for business surveys. These enable the compliance costs to business to be kept within limits agreed with Ministers. These plans and also regular reviews of major surveys involve independent outside observers. This means more attention should now be given to maintaining an effective overview and control of overall compliance costs. This should improve the priorities given to different measures to minimise burdens with less emphasis on looking at individual surveys in isolation as with the traditional methods of survey control.

3. USING ADMINISTRATIVE DATA

Partly as a result of decentralised organisation of government statistical work in the UK, with statisticians working closely with administrators in government departments, a range of administrative data is successfully used for statistical purposes. The main administrative sources used for the purpose of business statistics are the VAT register, the PAYE income tax system and company registrations which feed into the IDBR. Experience in the UK of using administrative data to help reduce statistical burdens on business and general observations arising from this is described in detail in a separate paper (see Machin 1997).

There are inevitably some limitations in the coverage and continuity of the IDBR but it does cover the businesses of interest for most surveys and is valuable generally for drawing samples. The strength of the IDBR lies in the integration of data from the various sources including supplementary survey information, especially on employment in local areas. It enhances the range of sub-national data available. However a drawback is the lack of a common reference number system for businesses. In linking up the separate details from different sources it is necessary to resort to name matching which is not an ideal method.

There are thus some limitations to the use of administrative data sources although further possibilities continue to be considered. For example in ONS we have been considering how to make better use of employment data from PAYE income tax records to improve the precision of short-term employment series and thus help to reduce compliance costs. While it does not appear to be possible to use the administrative data as a direct substitute for survey data, it seems that the efficiency of estimates from the surveys can be improved through use of the PAYE employment as a stratification and auxiliary variable.

4. ELECTRONIC DATA COLLECTION AND BUSINESS SOFTWARE

The ONS has used various forms of electronic data collection for business surveys, so far mostly on an experimental basis. The main focus is now on developing collection using specific reports included as separate modules in commercial accounting software packages. The use of EDI (Electronic Data Interchange) is well established for the collection of trade statistics (Intrastat) by Customs and Excise.

Full EDI replaces four elements in the process of supplying statistical data. It eliminates the need to:

- obtain a physical report from the business' system,
- transfer the relevant data to the statistical form,
- transmit or post the form back to the statistical agency,
- enter the data into the statistical system.

For the contributor it is the first two of these that provide the major savings with little benefit from replacing the postal system as the means of returning data. From the viewpoint of the statistical agency, however, it is replacing the second, third and fourth of these processes that would bring the greatest immediate savings. Early work in the UK thus concentrated on the method of transmission but it is now realised that the automatic extraction of data from existing company records should be the priority.

Very few businesses in the UK are so far able to send data by full EDI and few are willing to adapt systems just for the occasional few surveys. A major obstacle to the use of EDI is that it requires contributors and/or the statistical agency to invest in appropriate software and access to a communications network. Without sufficient incentives or the promise of significant savings of effort it is difficult for the contributor to justify such investment unless they are already using EDI for other business purposes.

Intrastat presently accounts for the bulk of statistical data collected using EDI in the UK. Nevertheless the volume of data keyed from paper remains significant and has increased. Initially the rejection rate of incoming EDI messages was high. One difficulty was that the system relied on commercially developed software and, as a result, over a hundred different packages were produced. There was no official certification of the software and thus no control over its quality. The emphasis was on allowing choice and reducing the burdens on business. The position has, however, now much improved through education of traders and the development of a system for the prompt identification and correction of errors. Various strategies have been deployed to make it easier for traders to use EDI for Intrastat rather than submit paper returns. Help has been provided in setting up systems, providing communication links and evaluating and approving software packages produced for Intrastat use.

ONS have had some success with electronic questionnaires which involve providing business with software, usually on floppy disks, but this is expensive. Once business has seen electronic questionnaires working well, this may help them to see the benefits of automatic extraction of statistical data from their own databases. However it seems unlikely that they will be universally used by all contributors.

It seems that the future of the collection of business data lies in making use of the information that business require to operate for their own purposes. While much administrative data can be captured from government sources and has a crucial role in the maintenance of business registers, the process can sometimes be too slow or inefficient to meet all requirements. ONS is therefore working with commercial software companies to develop modules in accounting software packages that meet ONS survey requirements using standard reports. The first software incorporating a module for ONS statistics is now due to be launched. The report module is very simple but it is hoped that sufficient businesses will participate in trials to identify problems and future enhancements.

There are three major advantages of this approach:

- there is no need to provide the software needed to provide the data,
- the data comes directly from the business internal systems with the resulting benefits to quality and compliance costs,
- as the recurrent burden is substantially reduced , it may be possible for the statistical office to collect more information more frequently than would be acceptable using conventional methods.

However, not all the information required for the statistics is available within a package and the information that is available may not meet the precise definitions required. Statisticians also need to be more active in anticipating changing requirements of users as the time taken to amend systems to collect different information will be much greater compared with using traditional paper forms. It is necessary to be involved with accounting standards bodies in the UK to influence the definitions used in commercial accounts. A more consistent and coherent strategy for data collection is also required, moving further away from the traditional piece-meal approach which may involve individual surveys asking for similar information in a slightly different way.

It is clear that a number of issues will need to be tackled:

- Dealing with the flexibility in UK accounting systems. The lack of a standard general ledger structure makes it difficult to develop software modules that would be effective without customisation for individual businesses. One possibility is that ONS might provide guidance to companies on how to produce what is needed from what they already have using existing report generators in software modules. It seems essential that ONS report modules that are developed allow businesses to control and change the way data are input. (Whilst this would provide business with an opportunity to make mistakes, this is no worse than the present position with paper forms.)
- Finding ways to make future initiatives commercially attractive to software houses and participating businesses, in particular to persuade accounting software companies to modify packages for the relatively few

users who get particular government survey forms. (The relatively low cost to business of providing statistical information makes them reluctant to invest in systems to support electronic statistical returns.)

- The need for an integrated approach to meet a variety of government requirements, not just for statistics, avoiding duplication of systems collecting similar information.
- The need to extend data collection to cover information, wider than that available from company accounts. Access to other systems, such as production, delivery and purchasing modules, payroll and personnel systems also need to be considered.
- So far, the reluctance of many businesses to deal with government electronically and the lack of a widely available, secure and reliable electronic means of communicating in the business environment.

5. THE WAY FORWARD

Much of the reduction in compliance costs identified by the Osmotherly review is coming from the streamlining of surveys, sampling more efficiently, reducing overlaps between surveys and the development of business software. The Osmotherly study concluded there was some scope for the greater use of data from administrative sources, but this was not seen as a major means of reducing burdens. However it seems significant further progress in the UK in the use of administrative data for the purposes of business statistics would require more timely data, greater co-ordination of definitions and classifications and further integration of the data sources. A common, or linked, reference numbering system for businesses would help considerably. But the opportunities under present arrangements seem limited. A key strategy in the UK is therefore to tap at source the administrative data that are used to operate the business.

Looking further ahead, there may be scope for more sharing of data required for other purposes but, as with the idea of using standard business software, this would not necessarily mean intercepting administrative data stored on other government departments' computers but making the most of information already available within businesses in a strongly co-ordinated manner to meet a range of administrative and statistical requirements. A wide view would need to be taken of the quality of data required for all users.

The UK Government intends to implement a strategy for sharing information from businesses (or individuals) needed by more than one government department where this is legally permissible (UK Government Green Paper 1996). There is no intention of merging all the sensitive information held on a business (or individual) into a single database. The idea is to create a new infrastructure to provide the link between the departments' information systems – which should remain separate- and individual businesses (or people). The full implications of this for statistics are yet to be assessed. There are clearly many important issues to be considered such as safeguards of confidentiality, suitable definitions and quality standards for shared data, mechanisms for sample checks *etc*. But such a development may present considerable opportunities for statisticians to make more use of information already available for other purposes both to enhance the range and quality of statistics and to produce them more efficiently with minimal burdens on business.

REFERENCES

Machin, A. (1997). Use of administrative data to reduce statistical burdens on business: Experience in the United Kindom. *Proceedings of the seminar on the use of administrative sources for statistical purposes*, Statistical Office of the European Communities.

Office for National Statistics (1996). Business Charter: A partnership with business.

Office for National Statistics (1997). Compliance Plan 1997-1999.

Perry, J. (1995). The Inter-Departmental Business Register. *Economic Trends*.

UK Government (1996). Statistical Surveys: Easing the Burden on Business, September.

UK Government (1996). Government Response to the Osmotherly Steering Group Report on Statistical Surveys, September.

UK Government Green Paper (1996). Government.direct – A prospectus for the Electronic Delivery of Government Services.

Walker, G. (1997). Collecting data using EDI: Experience in the United Kindom. *Proceedings of the seminar on the use of administrative sources for statistical purposes*, Statistical Office of the European Communities.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

SURVEY INTEGRATION: COLLECTING MORE DATA FROM FEWER RESPONDENTS

E. O'Brien and T. Wickwire[1]

ABSTRACT

In 1996, the National Agricultural Statistics Service (NASS) integrated two cross-sectional surveys, one environmental and one economic, to produce a new, multi-purpose data series of farm businesses. To balance the increased burden on individual respondents with client demands for a more comprehensive data set, NASS designed a single multi-mode, multi-phase survey. With the integrated survey approach, NASS minimized the burden that would have resulted from the uncoordinated data collection efforts of several USDA agencies, state departments of agriculture, and land grant universities. Per interview times decreased while numbers of usable reports increased. Data users have gained a data series with more analytic power.

KEY WORDS: Survey integration; Economic and environmental; Policy relevance.

1. IMPLEMENTATION OF THE INTEGRATED SURVEY

Integration of survey data is an implicit approach toward addressing the essential principles of a federal statistical agency. The principles, enumerated by Martin and Straf (1992), state that the federal statistical agency should produce data series which are policy relevant, are credible among data users, and earn the respect and trust of those who provide the data, the respondents. The integration of two major data series by NASS attempts to promote these principles.

Prior to survey integration, NASS conducted two similar but independent surveys of production practices and farm chemical use. The Cropping Practices Survey (CPS), was a multi-phase survey of crop yield estimates, production practices, fertilizer and chemical use. The Farm Costs and Returns Survey (FCRS), described the financial well being of farm businesses, farm households and specific farm enterprises such as corn or dairy production. The Cost of Production Survey (COP), a sub-sample of the FCRS, was used to develop commodity enterprise budgets from production practices measures similar to but less comprehensive than CPS items.

Integration involves several levels: integration of concepts, data inputs, data processing and data outputs (Colledge 1990). In discussing outcomes, therefore, it is useful to understand the processes behind the new survey design by evaluating these dimensions.

In 1990 NASS began publishing environmental data, specifically chemical use data, from the CPS. The CPS survey design had several disadvantages.

- Coverage was poor. Chemical applications by product and area treated were published on just five crops: corn, soybeans, wheat, cotton and potatoes. Though valuable in describing that fixed set of crops, too little data were available on U.S. agriculture as a whole. Rare commodities, such as fruits, vegetables and specialty crops like tobacco were ill-suited to the CPS area frame sample design. Farmers were at risk of losing chemicals vital to their industry because independent, unbiased data were simply unavailable to federal regulators.
- It could not meet the test of policy relevance. Chemical use data were collected in a multi-phase survey where other phases were devoted to production measures. Production data added limited analytical value to chemical use data for researchers challenged to evaluate the economic impact of the department's emerging environmental policies.
- Coordination and integration with surveys outside NASS was not possible. Federal initiatives and funds fuelled the development of a national environmental policy. Reaching for an objective source of information upon which to evaluate a national policy, federal agencies, state governments, and land grant universities turned to surveys. A framework began to evolve but with insufficient integration of concepts, measures and resources. Clearly the respondents, members of a dwindling and over-surveyed population, were not well served.

The environmental data series needed a more flexible design to increase crop coverage, incorporate new subject areas, economic measures, and clients. Fundamentally, the survey had to grow from producing a descriptive data series to supporting a complex analytic and research framework

1.1 ARMS Implementation: The Integrated Survey Design

Survey integration had these main goals: to develop a design which could better address emerging policy questions, to produce a more complex data set while con-

[1] Eileen M. O'Brien and Trent Wickwire, U.S. Department of Agriculture, NASS, 1400 Independence Ave., Room 1462, Washington, D.C. 20250, U.S.A.

solidating data collection funds of several clients, and to minimize respondent burden while requesting more data of individual respondents. The integrated survey, the Agricultural Resource Management Study (ARMS), kept the production practices and environmental components of the CPS and linked them with the enterprise and whole farm financial pieces of the FCRS/COP.

Developing the ARMS proved more arduous than the simple union of the CPS and FCRS/COP designs. At what level would the data be integrated? CPS data had been collected on a PPS selected field. FCRS/COP data had been collected at the commodity enterprise and whole farm level. Which sample frame would be used? The CPS used an area frame sample with multiple personal interviews while the FCRS was a list dominant design using a single, lengthy personal interview. The options were to merge the CPS into the FCRS, merge the FCRS into the CPS, or stay the course and field two independent surveys as the large, uncoordinated efforts outside NASS continued to grow. A step toward burden reduction was not obvious in any of these scenarios.

The cornerstone of the ARMS design was the screening phase. Besides the ARMS sample, Phase I screening yielded sample for seven customer surveys outside NASS. Phase I served several purposes: to determine which records were in scope and had targeted commodities; to remove duplicates, refusals and inaccessibles; and to identify where subsampling of complex operations was necessary. Ultimately, screening improved the quality of cost of production estimates by increasing the number of usable reports. By removing out of scope records, cost per usable field interview was expected to fall. Each year as department economists update other cost of production budgets, NASS will rotate in different targeted commodities, not possible under the old CPS design. Phase I screening may be adapted to increase sample for special data collection efforts for state governments, land grant universities and other USDA agencies.

Besides accommodating additional crops, the ARMS design can accommodate additional phases or modules of questions. Adding phases permits complex analysis across the level at which production, financial, management or household decisions are made. For example, some economists will estimate the cost of producing a corn crop in 1996 from production practices information on a randomly selected field in Phase II, and enterprise and farm level expenses in Phase III. Another group of economists will evaluate the costs and benefits of different conservation practices and programs; how farmers manage risk under the new farm program; the changing structure of farm income, food prices and the environment; and barriers to adoption of environmentally friendly production practices.

1.2 Arms and Respondent Burden Reduction

Under the old design a farm might be chosen for the CPS then for the FCRS or FCRS/COP. NASS minimized selection for more than one survey of similar content by simultaneously classifying sample units to be eligible for the ARMS or its companion surveys. A formal sample selection procedure insures that the fewest number of respondents be in the ARMS and other equally burdensome surveys. Through survey integration, survey concepts are harmonized and duplication of questions is minimized.

In the old COP design all data were collected in one 46 page Spring interview averaging 2 to 2.5 hours. In the ARMS design (24-28 pages), Phase II interviews averaged less than one hour while few Phase III interviews exceeded 2 hours. To address burden, eighty pretests and discussions with department economists revealed where demands could be lessened with little or no loss in data utility or integrity. Survey methodologists evaluated which items were best reported in field-level Fall interviews versus enterprise level Spring interviews based on reliance on memory versus records. Since fewer records are kept on production practices data, these data are collected in the Fall to minimize recall error while whole farm financial questions remain in the Spring interviews so complete records may be used (Willimack and O'Brien 1995).

With production practices data, respondent's are asked to recall specific events. In the COP some recordless items were encoded in memory 14 months prior to the interview. Respondents were expected to report for the whole commodity enterprise. The ARMS design reduces burden by limiting these data items to a selected field (Table 1) using a more focussed question protocol. To foster internal consistency of reported data, variable costs associated with field activities were also moved from the Phase III enterprise to Phase II field level questionnaire. Reported costs are now directly related to specific field activities.

Table 1
Level of Data Collection, Chemical Use and Production Practices

	Old FCRS/COP	Old CPS	New ARMS
Fertilizer, Chemical Use	Enterprise	Field	Field
Capital Equipment Use	Enterprise	Field	Field
Field Operations	Enterprise	Field	Field

Where records are maintained, the new design attempts to borrow from existing motivation to keep records (O'Brien 1997). Because tax requirements provide a stronger motivation than surveys to draw financial records together, most financial information is collected in Phase III Spring interviews (Table 2). The survey data collection schedule goes beyond the April 15 tax due date to help respondents report financial data from complete records.

Table 2
Level of Data Collection, EXPENSES

	Old FCRS/COP	Old CPS	New ARMS	
			Variable	Fixed
Direct Production	Enterprise & Farm Level	–	Field	Enterprise & Farm Level
Labour Expenses	Enterprise & Farm Level	–	Field	Enterprise & Farm Level

Final datasets were delivered to users in July 1997. Though too soon to analyse data quality, a review of the process is possible. In general, usable reports have increased (Table 3). In-person response rates are comparable to the CPS versus the less successful FCRS. Interview lengths did not meet 60 minute targets in all cases, but sizeable decreases were achieved.

Table 3
Number of Completed Reports, Old v. New Design

	Corn for Grain		Flue-cured Tobacco		Beef Cow/Calf	
	FCRS 1991	ARMS 1996	FCRS 1991	ARMS 1996	FCRS 1990	ARMS 1996
Sample Size.....	2,034	2,089	611	480	3,000	1,720
Completed	708	1,393	242	316	819	1,279
w/ commodity	(35%)	(67%)	(40%)	(66%)	(27%)	(74%)
w/o commodity..	272	–	117	–	882	
	(13%)		(19%)		(29%)	

Commodity specific ARMS versions for Corn, Flue-cured Tobacco and Beef Cow/Calf production can be compared against previous FCRS/COP surveys. While the field sample size was nearly identical for corn, usables clearly increased from Phase I screening. Each commodity, but especially Beef, showed a more than 20 point increase in completions with the targeted commodity. Essentially the greatest gains are for the rarest commodities.

Though interview times surpassed the 60 minute threshold, per interview time was dramatically reduced (Table 4). Special questionnaire supplements may have inflated the ARMS Phase III interview times by 10-20 minutes. Content in Phase II Flue-cured tobacco questionnaires was not sufficiently controlled. In future years, interview times will have to be better managed across both phases to reach the 60 minute per interview goal.

Table 4
Questionnaire Interview Times, Old v. New Design

	Corn for Grain	Flue-cured Tobacco	Beef Cow/Calf
		(minutes)	
FCRS/COP	123	145	130
ARMS II, 1996	63	93	53
ARMS III, 1996	89	83	79
TOTAL	**152**	**176**	**132**

Overall response rates have not appreciably increased but rather inaccessibles and refusals have shifted from a more costly field interview mode to the less costly telephone pre-screening mode (Table 5; Rutz and O'Brien 1997). Several survey characteristics explain these low rates. Unlike establishment surveys outside agriculture, ARMS uses a probability sample rather than a quota sample of willing businesses; survey cooperation is not mandatory as is customary in other business surveys; there are no refusal conversion attempts; the survey topic is both very sensitive and difficult; and finally, the public reporting burden on this survey population is very high. Obtaining cooperation from large farm businesses is a persistent problem in both designs.

Table 5
Response Rates by Version, Old v. New Design

	1995 FCRS	1996 ARMS II
	(percent)	*(percent)*
Whole Farm Financial	**55**	**56**
Commodity Enterprise Level	57	72
TOTAL	**55**	**59**

3. DISCUSSION

Those conducting demographic surveys sample from a growing population while NASS and others who conduct business surveys, must sample from a dwindling population. While NASS conducts 800 surveys a year, there are fewer farms to report and their average size is growing.

About 25 percent of U.S. farms produce 75 percent of the nation's food and fibre. Drawing a sample which accurately represents U.S. agriculture necessarily increases the reporting burden on larger, commercial farms. State governments, land grant universities, and many private organizations also make survey demands on farm businesses. Besides survey reporting burden, administrative and regulatory record keeping requirements are expanding. NASS and its customers will be challenged to maintain the respect and trust of these respondents on voluntary surveys while acknowledging the larger public reporting burden farms face. Survey integration, more accommodating reporting formats and schedules are just a few tools.

Where survey integration has taken place, much has been written about analytic synergy, filling data gaps, and cost sharing under increasing budgetary stress (Cohen 1996; Colledge 1992; Madans and Hunter 1996). Few have addressed the effect of making repeated, complex survey requests where the respondent pool is shrinking. Too many requests may accelerate the decline in response rates for voluntary surveys. More research is needed to understand which methodological innovations may reverse that trend, and which efforts may have a lasting effect.

3.1 The Future of Survey Integration

There are many advantages in the ARMS design which serve data users. Adapting to new data needs proved challenging in the production environment. Consolidation of data collection funds demanded team building among stakeholders with very different research interests. The effort in managing concepts, measures, the survey process and budgets was underestimated. Data summarization was considerably more complicated than anticipated. Ultimately, a more useful data series has begun and further gains in addressing respondent burden show promise.

Future research efforts should examine whether there is differential nonresponse by mode. Would certain respondents refuse Phase I screening by telephone but consent to subsequent in-person interviews? Special topics could be modularized and conducted on a subsample to move toward the 60 minute per interview goal. Renewed emphasis should be placed on training interviewers to effectively address concerns about the survey topic and respondent reluctance. Measurement error identified in previous research should be addressed while respondent's growing reluctance to participate raises concerns about nonresponse bias of survey estimates.

REFERENCES

Colledge, M. (1990). Integration of economic data: benefits and problems. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada, October 1990, 51-63.

Cohen, S. (1996). The Redesign of the medical expenditures panel survey: a component of the DHHS survey integration plan. *Statistical Policy Working Paper 26*, Washington, D.C.: Statistical Policy Office, U.S. Office of Management and Budget.

Madans, J., and Hunter, E.L. (1996). Implications for NCHS Data Systems. *Statistical Policy Working Paper 26*, Washington, D.C.: Statistical Policy Office, U.S. Office of Management and Budget.

Martin, M.E., and Straf, M.L. (1992). *Principles and Practices for a Federal Statistical Agency*. National Academy Press, Washington, D.C.

O'Brien, E. (1997). Redesigning Economic Surveys of Establishments, presented at the 52nd Annual Conference of the American Association of Public Opinion Research, Norfolk, VA.

Rutz, J., and O'Brien, E. (1997). Survey Integration: The Complete Data Set. Presented at the International Field Directors and Field Technologies Conference, Norfolk.

Willimack, D., and O'Brien, E. (1995). Record Keeping Systems as a Data Collection Mode for FCRS: Status Report. USDA-NASS SMD Briefing.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

INTEGRATING SURVEYS

J. Winkels and F. Kerssemakers[1]

ABSTRACT

Several developments brought Statistics Netherlands to go a step beyond harmonization of survey variables. These developments especially have to do with the demand for statistics on living conditions, initiatives to improve response in social surveys and further development of information technology. An overview is given of the most important reasons that lay the new design of a system of social surveys.

KEY WORDS: Integration; CAPI; Surveys; Harmonization; BLAISE.

1. INTRODUCTION

As in other European countries Statistics Netherlands has designed several surveys to gather information on (aspects) of living conditions. These surveys not only vary in content, but also in the sample frame used, sample size, response, interview length, number of persons to be interviewed, use of proxy respondents, type of data editing, weighing method, *etc.* Some of these surveys were revised in the beginning of the nineties when Statistics Netherlands introduced computer assisted interviewing (CAPI). These revisions focused both on the harmonization of questions used for demographic and socio-economic classifications and on the translation into CAPI of questionnaires that were originally developed for being asked with paper and pencil (PAPI).

Several developments brought Statistics Netherlands to go a step beyond harmonization of survey variables. These developments especially have to do with the demand for statistics on living conditions, initiatives to improve response in social surveys and further development of information technology. In the next section an overview is given of the most important reasons that lay the new design of a system of social surveys. The name of the system, POLS, refers both to the Dutch abbreviation for a Continuous set of Surveys on the Quality of Life and the Dutch word for pulse ('to keep a finger on the pulse'). The POLS-design is presented in section 3. Sections 4 focuses on the integration from a Blaise perspective.

2. THE INTEGRATION OF SOCIAL SURVEYS

At this moment Statistics Netherlands has started the redesign of its data collection on persons and households. This redesign is, because of the enormous complexity and impact, structured along some stages. In the first stage the socio-cultural surveys are redesigned. From 1997 onwards the seven surveys in the fields of health, crime and security, political and social attitudes, the general social survey (with a special survey on the youth and the elderly) and housing are being combined. However, the long term developments that steer the redesign of these surveys will also influence the redesign of other surveys inside and (as we see it) outside Statistics Netherlands in the near future: new demands, the fight against non-response and new technology to link data.

2.1 New Demands

Rather than being concerned with the distribution of aggregate variables over groups of statistical units, social research frequently focuses on relations between variables at the micro-level. This focus on micro-relations has been the driving force behind the development of ever more comprehensive household surveys. These surveys aim to cover as many variables as possible, so that all social relations can be analysed for a comprehensive data set. However, there is a growing awareness of the limits to this approach. The response burden on the sample households becomes too heavy if too many variables are covered in a single survey. This burden forms a serious constraint on statistical agencies to meet all the demands. The type of statistical information needed by policy makers also seems to change. The demand is no longer directed to rather one-dimensional statistics about the well known themes of the political agenda (health, crime, income). The demand for information directed to the more complex (causal) relations between these themes has increased. This demand reflects the policy questions arising from societal developments such as poverty, social exclusion and deprivation. These societal problems often converge in problem areas, like inner cities. Especially these developments need to be monitored by statisticians to fulfil the politicians needs. To state it in 'variable language': not the univariate or bivariate distributions of variables but the multivariate relations between sets of variables is being asked for.

2.2 The Fight Against Non-Response

Non-response in household surveys is a severe problem. Compared with other countries The Netherlands are not doing very well. Within the domain of the socio-cultural

[1] Jeroen Winkels and Frans Kerssemakers, Statistics Netherlands (CBS); e-mail: JWNS@CBS.NL.

surveys response varies between 50% and 60%. This level of response probably causes serious bias that can not always be neutralized by smart methods of weighing. With respect to the four-yearly National Election Study, for example, there is a small debate in the Netherlands what the figures from this survey really tell us about political alienation when so many people do not participate. Statistics Netherlands has taken several initiatives to reduce non-response. A lot of these initiatives have to do with the way surveys are presented to the public (*e.g.*, introduction letters), the monitoring of interviewers and the optimization of using different data collection methods. These initiatives are always taken under the restriction that the overall respondent burden should be as low as possible.

2.3 New Technology to Link Data

Other reasons for changing the social surveys of Statistics Netherlands stem from information technology developments. There are basically three methods to link micro data from different sources. The first is exact matching. In this case micro data from different sources on the same individual are linked. The second way of linkage is synthetic matching. In this case micro data from different sources on the same group of individuals (for example age group) are linked. A third method of linking micro data is to redesign the sources in such a way that there will be a brief joint questionnaire for a large sample that provides succinct information on core variables from all original surveys; and more in depth questionnaires on the separate areas for smaller subsamples. This way a core micro data set is obtained directly (the equivalent of exact matching) and an in depth synthetic data set can be created by combining the in depth data from the various subsamples using the joint core as a synthetic matching key. Because the variables in the core are associated with the in depth variables of the subsamples, this approach picks up relations more easily than synthetically matched data sets that rely on just demographic variables for the matching. POLS has been based on this third method.

2.4 Restrictions on the Integration of Surveys in POLS

The most important restrictions that steered the design of integrating the separate social surveys within the socio-cultural domain were:

(1) Minimization of the burden on respondents. This condition can be viewed from a micro and from a macro perspective. 'Micro' means that the mean interview-time within a household should not exceed 45 minutes. 'Macro' means that the total interview-time of all the modules should be reduced by POLS. This restriction has also lead to efforts to combine separate surveys from other agencies with specific POLS-modules.

(2) A person-based sample frame. Statistics Netherlands now has the possibility to take samples from a file that is based on the fully automated population register in the Netherlands. In the first place the fieldwork department can profit from this information to combat non-response: interviewers know beforehand who is going to be interviewed and some characteristics of the persons who refuse are known. But data-users can also profit from this information, because oversampling is rather easy, because some information on the non-responding persons is given and because this register will (in the future) be combined with other registers.

(3) Enough cases to allow description of relatively small subgroups. This condition could only be met if all the socio-cultural surveys should join. However, the estimated net response of 36,000 households (1997) forces Statistics Netherlands to make work of the extension with other surveys to find the smallest ones of the target groups for policy makers.

(4) No proxy interviewing (for persons who are older than eleven years), unless the empirical relation with other characteristics is well known. This restriction has to do with the efforts to both improve quality of the data (proxy information is in general less reliable) and to decrease the burden on respondents.

(5) Flexibility in adding modules on certain topics during certain periods or in oversampling specific population groups within a separate survey.

3. THE DESIGN OF AN INTEGRATED SYSTEM OF SOCIAL SURVEYS (POLS)

Shell-Structure

The design of POLS is based on a shell-structure. In principle there is no limit to the number of shells, so panel studies can be included. Every shell has its own characteristics (see following Figure). A survey always consists of the joint questionaire (shell 1 and shell 2) and one or more modules from shell 3. However, the order of the questions during the interview can be different, because of the interviewing process. For example, the questions on income that make part of the joint questionnaire are always asked at the end of the interview. The joint questionnaire is designed under the restriction that both telephone and face-to-face interviewing should be possible. The CATI-part will be used to reach certain types of non-respondents from the first fieldwork stage. At the end of this telephone interview the respondents are asked if they are willing to join in a face-to-face interview (to gather information being asked in shell 3).

Shell 1 (total sample)

Joint questionnaire part 1: harmonized classification variables

This part contains all the questions to be asked on every person in the sample. The questions use the harmonized classification variables, both the demographic ones (age, sex, marital status, nationality, place of birth) and the socio-economic ones (education, socio-economic position, household income). For the future it is planned to collect this information as much as possible via registrations. Apart from the future use of register information, this part of the questionnaire will only change if the concepts and definitions of the harmonized questions are changed.

Shell 2 (total sample)

Joint questionnaire part 2: core questions in the socio-cultural domain

This shell contains the core questions of the socio-cultural surveys. The reason to name it shell 2 (the interviewer off course does not see these terms) is that it is foreseen that in a later stage core variables from other surveys will be added. This shell is therefore less stable in content than shell 1. Because the questions in this shell are used for the total sample, they allow the quarterly publication of important indicators and the annual publication on a low regional level.

Joint questionnaire part 3: screening questions in the socio-cultural domain

This part varies from year to year. Although the large sample allows finding a lot of special groups, including a lot of screening questions in the joint questionnaire is not possible. In 1997 the screening (of 36,000 persons) is for accidents and injuries. The information gathered will be used to ask a specific subsample to join in a follow-up study (shell 4, see below).

Joint questionnaire part 4: Variable questions

To meet the fifth restriction (see above) POLS has implemented the principle that a small part of the joint questionnaire should be reserved for actual themes or themes that require a large sample. This part can vary twice a year. In 1997 it is used for questions on victimization.

Shell 3: Socio-cultural modules (subsamples)

In this shell theme-specific questions are planned. However themes are combined in a way that more-dimensional indicators (as described above) can be calculated. This calculation will partly be based on the idea of consistent weighting. The contents of this shell change regularly. Data on some topics have to be collected once or twice a year during a certain period. Some other topics only need to be analysed once every three or four years. Within this shell also screening questions (on a subsample) are included to trace persons for a follow-up study. The following modules, for which non-overlapping samples are used, will be implemented in 1997. The figures between brackets refer to response and age categories.

Health (N = 10,000, age 0+). This module uses two modes of data collection: CAPI and PAPI. The paper and pencil mode is primarily chosen because of the sensitive questions on health, for example the 'burnt out' syndrome. The module is continuously being asked (January-December).

Justice and Environment (N = 7,500, age 12+). Here again CAPI and PAPI are both used. The PAPI is chosen because of the questions about time use (a so called 'yesterday interview'). The module is continuously being asked.

Justice and Participation (N = 5,000, age 12+). This module only uses CAPI. Part of the questions are identical to the last module in order to fulfil the requirement of having 12,500 responses about themes like victimization, crime prevention *etc*. The module is continuously being asked.

Youngsters (N = 4,000, age 12-30). This module also only uses CAPI, but in a slightly different way. Because of the interviewed population (under 30 years) it is possible to ask the respondents to complete parts of the questionnaire on a notebook computer by themselves. The answers to these questions (about sex, drugs and crime) will be of a better quality if self-completion is chosen. The module is scheduled three times in a decade and will be asked from March 1997 till December 1997.

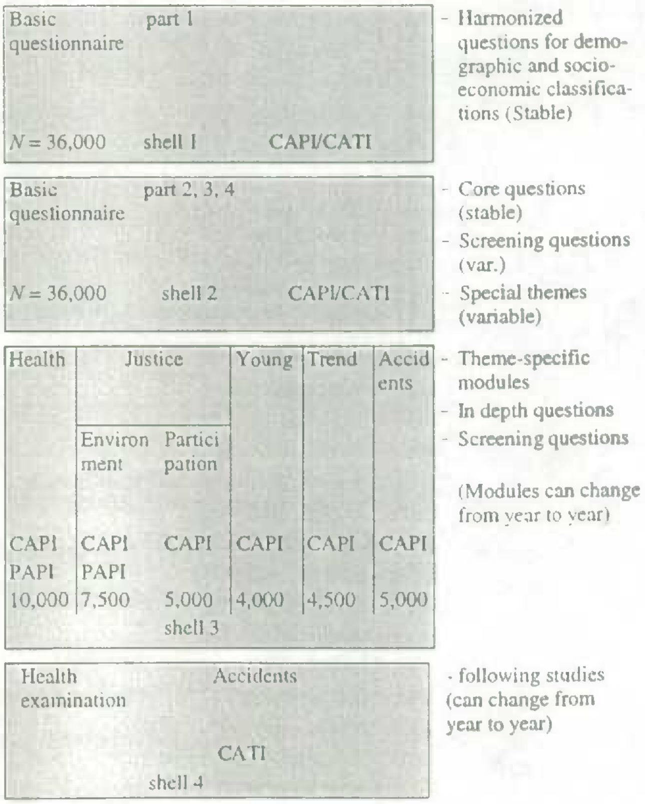
Trends (N = 4,500, age 18+). This module only uses CAPI and primarily consists of trend questions that have been asked in surveys since 1974, and questions on request of the Social and Cultural Planning Office. The module is scheduled three times in a decade and will be asked from March 1997 till December 1997.

Accidents (N = 5,000, age 0+). The survey on accidents is part of a new research project of Statistics Netherlands that has been formulated on request of the Ministry of Health. The other part of the data-need of this project has been formulated within the joint questionnaire (screening questions) and a follow-up study in shell 4. This module uses CAPI and PAPI. The PAPI-part is the same as described above: a yesterday interview to measure time use. Here it is only to be used on Friday however, in order to supply additional cases for getting a uniform distribution of net responses over the days of the week.

3.5 Shell 4: Follow-up studies (subsamples)

The possibilities both in the joint questionnaire and in the specific modules to screen target groups are used for follow-ups. In 1997 one follow-up study is foreseen on accidents and injuries (CATI-interview of about 15 minutes)

An overview of the Integrated System of Surveys (POLS) in the year 1997.
(Numbers refer to persons sampled from Population Register.)



As the mean interview-time for the joint questionaire is fixed at 15 minutes, shell 3 may generally take another 30 minutes.

In 1997 a reduction of the total (macro) interview-time of 20% (*i.e.*, 5,000 interview hours) will be achieved thanks to the introduction of POLS.

4. INTEGRATING SURVEYS FROM A BLAISE PERSPECTIVE

To keep things manageable the in principle indefinite number of subject-matter questionnaires should be independent from each other as much as possible. The integration concept rests entirely on the common joint questionnaire. The latter could be duplicated in different datamodels, one for every subject-specific questionnaire. These datamodels could then be put under the umbrella of a single case management system and be handled by Maniplus as a single survey. But especially the joint questionnaire is meant to be relatively stable and independent from the variety of subject-oriented wishes and developments. And if the datamodels are to be recognizable entities linked to a particular subject, changing the joint module should leave them unchanged, if possible. Subject-matter specialists who are responsible for a particular datamodel should not be bothered with things they did not initiate and that may be irrelevant to them. Therefore it was decided to have a separate datamodel for the joint questionnaire on the integrative level next to a set of datamodels for which, as before, independent and non-overlapping samples of persons are drawn.

First, the joint questionnaire is asked. Included are also themes that require a large sample (*e.g.*, victimization or accidence rates). After finishing this datamodel a Manipula-setup takes care of writing the data that may possibly be needed in anyone of the shell 3-datamodels in a separate 'extemal' datamodel. Through a dialogue box in Maniplus the interviewer can now open the particular datamodel to which the target person was assigned. After concluding this part of the interview some questions still have to be asked about the household of the target person. Here data are collected from the person with the highest income, such as income itself, educational and occupation. As this may involve a shift of respondent these questions, which actually belong to the joint questionnaire, are asked near the end of the interview. They are put in a separate datamodel, which is only auxiliary because the data is sub-sequently added by Manipula (block moving) to the files of the original datamodel. Thus, the latter will finally contain all data from the joint questionnaire. For a particular case only the data from the first model and from the applied shell 3 are sent back by telephone. So, for the actual questionnaires POLS uses a 1-n-1 structure of datamodels. The two constants represent the data-models for the generally applied joint questionnaire. The variable n represents the datamodels for the different shells 3 from which one is chosen per interview. Afterwards extra data-models can be used, as they are, for follow-up studies among screened cases (*i.e.*, shell 4).

The interviewer can always return to preceding datamodels for changing already given answers. However, as this may effect routing, checks and computations in subsequent models, Maniplus forces the interviewer to open these models again so that potential changes will be processed automatically. This could be inconvenient if occurring frequently. In practice it can be coped with by choosing the right, relatively independent modules.

SESSION C-9

Data Analysis: Where We're at and What's New?

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

DESIGN AND ANALYSIS OF EXPERIMENTS EMBEDDED IN COMPLEX SAMPLE SURVEYS

J.A. Van den Brakel and R.H. Renssen[1]

ABSTRACT

This paper focusses on the design and analysis of large scaled field experiments embedded in ongoing surveys. We will discuss how parallels between randomized experiments and random survey samples can support the design as well as the analysis of field experiments embedded in ongoing surveys. Generally, the objective of embedded field experiments is to draw inference on finite population parameters. To this end the analysis has to include the probability structure imposed by the chosen sampling design of the survey as well as the randomization of the chosen experimental design used to assign the experimental units to the different treatments. We worked out such a design-based method for the analysis of the completely randomized design with the two-sample problem as a special case.

KEY WORDS: Completely randomized designs; Design based inference; Embedded field experiments; Model based inference; Randomized block designs; Two sample problem.

1. INTRODUCTION

At Statistics Netherlands, several large scale field experiments embedded in ongoing surveys have been conducted in order to test or quantify the effect of different design parameters of a survey on response rates or estimates of finite population parameters of a sample survey. In Van den Brakel and Renssen (1998a) a series of embedded field experiments conducted at Statistics Netherlands are described. Embedded experiments are particularly appropriate if interest is focussed on the quantification of the effect of alternative survey methodologies on estimates of finite population parameters of the sample survey. In these experiments the sample of an ongoing survey is randomly divided into one relative large subsample and one or more smaller subsamples. The large subsample is assigned to the regular survey and serves, besides publication purposes, as the standard methodology or control group. The other subsamples are assigned to the alternative treatments and are conducted parallel with the regular survey.

Fienberg and Tanur (1987, 1988, 1989) emphasized that the statistical methodology used in randomized experiments and randomized sampling is essentially the same. Parallels between the concepts of randomized experiments and random sampling can be exploited in the design of embedded experiments to improve the accuracy of estimated treatment effects and to draw correct conclusions about the observed effects (*i.e.*, the internal validity of the experiment). For example, because respondents interviewed by the same interviewer produce more homogeneous answers than respondents interviewed by different interviewers, it is efficient to apply local control for interviewers by means of randomized block designs (RBD). The concepts of randomized block designs and stratification are essentially the same. If an experiment is embedded in a stratified sampling design, consequently, this parallel leads naturally to a RBD with strata as block variables. If in two stage sampling designs the sampling units from the same primary sampling units (PSU's) are more homogeneous than sampling units from different PSU's, it will be efficient to assign the *k* treatments randomly to the secondary sampling units within each PSU or cluster. This type of randomization naturally leads to a split plot design where the PSU's correspond with the whole plots or a RBD where the PSU's are modelled as random block variables. A more detailed discussion of how to take advantage of the structure of the sampling design in the randomization of the experimental units over the different treatments can be found in Fienberg and Tanur (1987, 1988, 1989) and Van den Brakel and Renssen (1996a, 1998a). In summary, the sampling design of the survey forms a prior framework for the design of an embedded experiment. In the following sections we propose a design based approach for the analysis of embedded field experiments.

2. ANALYSIS OF EMBEDDED FIELD EXPERIMENTS

In embedded experiments experimental units are selected by some complex sampling design from a finite population. Statistical methods traditionally used in the analysis of experimental designs are model dependent and require IID observations; however, the stochastic assumptions underlying these techniques do not reflect the complexity which is usually exhibited by the applied sampling design and the finite survey population from which the experimental units have been drawn (Skinner *et al.* 1989, Ch.1). In these cases the assumption that the observations are IID is usually not tenable.

[1] Jan A. Van den Brakel and Robbert H. Renssen, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands. The authors wish to thank Prof. Stephen E. Fienberg for reading of drafts of this paper.

Fienberg and Tanur (1987, 1988, 1989) advocated a model based approach for the analysis of embedded field experiments. The internal validity is ensured by the application of fundamentals as randomization and local control on sampling structures like strata, clusters or interviewers in the design as well as in the analysis of the embedded field experiment. The external validity is achieved by incorporating certain local control variables like interviewers or clusters as random components in a mixed model analysis. Fienberg and Tanur (1988) showed, using statistical likelihood theory, that the weighting of the applied sampling design can be ignored in the analysis if the selection of the sampling units depends only on prior variables which are conditioned on in the statistical model and are independent of the target variables. If the experiment is analyzed under the assumption of IID observations, then the analysis is performed conditionally on the drawn sample and inferences are made about the parameters of some hypothetical superpopulation model and not about the finite population from which the sample is drawn.

The purpose of most embedded field experiments is to test or quantify the effect of alternative treatments on estimates of finite population parameters. The disadvantages of the model based approach is that the inference concerns model parameters from some superpopulation but not the estimates of the parameters of the finite survey population, even if the external validity is guaranteed by using random or mixed models. Furthermore the validity of the inference depends on model assumptions. Therefore a design based approach is emphasized in this paper. In embedded field experiments a large number of experimental units are selected from a finite population by means of a random sampling design. As a result, it becomes possible to draw inferences about finite population parameters that do depend on a probability structure imposed by the design of the survey and not on model parameters from a superpopulation that depends on an assumed probability distribution. To this end, the analysis should be based on the estimates of finite population parameters. From the objective of the experiment, sensible hypotheses about these finite population parameters are formulated and efficient test statistics are constructed. Statistical methods from sampling theory can be used in constructing such test statistics, which take into account that experimental units are selected from a finite population by some complex sampling design with possibly unequal inclusion probabilities and/or clustering. In the next section such a method for the analysis of completely randomized designs is proposed to illustrate the possibility of developing a design based method for embedded experiments.

3. DESIGN BASED APPROACH FOR EMBEDDED COMPLETELY RANDOMIZED DESIGNS

Consider an embedded field experiment designed to compare the impact of $K - 1$ alternative survey methodologies with respect to a standard survey methodology on a target parameter of a survey. Let $\bar{Y} = (\bar{Y}_1, \bar{Y}_2, ..., \bar{Y}_K)^t$ denote a vector of order K with each element the population mean of a target parameter measured by using treatment k. The objective of the experiment is to investigate whether there is a significant difference between the parameters $\bar{Y}_k, (k = 1, 2, ..., K)$. From this objective the following hypothesis can be derived:

$$H_0: \bar{Y}_1 = \bar{Y}_2 = ... = \bar{Y}_K$$

$$H_1: \bar{Y}_k \neq \bar{Y}_{k'}, \text{ for at least one pair,}$$
$$(k, k' = 1, 2, ..., K \text{ and } k \neq k'). \qquad (1)$$

Clearly, this hypothesis only concerns finite population parameters. To test this hypothesis, a sample s of size n is drawn from the target population U of size N by some complex sampling design with first order inclusion expectations π_i for sampling unit i and second order inclusion expectations π_{ij} for sampling units i and j. Regardless of the structure of the sampling design, the sample s is randomly divided into K subsamples s_k of sizes n_k. This experimental design is called a completely randomized design (CRD). The subsamples are not necessarily of equal size. The units of subsample s_k are assigned to one of the K treatments. Variable y_{ik} is defined as the response of experimental unit i assigned to treatment k.

To draw inference about finite population parameters, the analysis should explicitly take into account the probability structure of the applied complex sampling design used to draw sample s (established by the first and second order inclusion expectations π_i and π_{ij}) as well as the randomization mechanism of the experimental design applied to divide sample s into K subsamples. To this end it is proposed to test hypothesis (1) with the Wald test. Hypothesis (1) can also be written as $\boldsymbol{C}\bar{\boldsymbol{Y}} = \boldsymbol{0}$, where $\boldsymbol{0}$ is a vector of zeros of order $K - 1$ and $\boldsymbol{C}$ the $((K-1) \times K)$ matrix with contrasts, for example $\boldsymbol{C} = (\boldsymbol{j} \mid - \mathbf{I})$, where $\boldsymbol{j}$ denotes a vector of order $K - 1$ with each element one and $\boldsymbol{I}$ denotes the identity matrix of order $K - 1$.

In order to include the complexity of the sampling design into the analysis of the CRD, it is proposed to use a design unbiased estimator for the estimation of $\bar{Y}$ and its variance-covariance matrix. Let $\hat{\bar{Y}}$ denote a design unbiased estimator for $\bar{Y}$, $\sum$ the variance-covariance matrix of $\bar{Y}$ and $\hat{\sum}$ a design unbiased estimator for $\sum$. Hypothesis (1) can be tested against the alternative hypothesis that at least one pair is significantly different by means of the Wald statistic:

$$W = \hat{\bar{\boldsymbol{Y}}}{}^t \boldsymbol{C}^t (\boldsymbol{C} \hat{\sum} \boldsymbol{C}^t)^{-1} \boldsymbol{C} \hat{\bar{\boldsymbol{Y}}} \, . \qquad (2)$$

Due to the fact that a design unbiased estimator is applied to the estimation of the population parameters and its variance-covariance matrix, a statistical test that takes into account the complexity of the sampling design is obtained. First the Horvitz-Thompson estimator is used as a design unbiased estimator and then results are given for the generalized regression estimator.

The first order inclusion expectation for the sampling units in subsample s_k are $(n_k/n)\pi_i$ (Van den Brakel and

Renssen 1996c). It follows that the Horvitz-Thompson estimator for $\bar{Y}_k$ based on subsample s_k is

$$\hat{\bar{Y}}_{\pi_k} = \frac{n}{Nn_k}\sum_{i\in s_k}\frac{y_{ik}}{\pi_i}, \quad k = 1, 2, ..., K. \tag{3}$$

We propose vector $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{\pi}} = (\hat{\bar{Y}}_{\pi_1}, \hat{\bar{Y}}_{\pi_2}, ..., \hat{\bar{Y}}_{\pi_K})^t$ as a design unbiased estimator for $\bar{\boldsymbol{Y}}$ in test statistic (2). Because the estimates $\hat{\bar{Y}}_{\pi_k}$ are based on K interpenetrating subsamples drawn from a finite population, they are not independent. In Van den Brakel and Renssen (1996c) an expression for the variance covariance matrix $\sum$, taking into account the dependency between the elements of $\hat{\bar{\boldsymbol{Y}}}_{\boldsymbol{\pi}}$, is derived. To derive an unbiased estimator for $\sum$, for each i, a vector $y_i = (y_{i1}, y_{i2}, ..., y_{iK})^t$ containing the observations of all K treatments for experimental unit i is required. These paired observations are not available because the sampling units are assigned to exactly one of the K treatments. However, an approximately unbiased estimator for $\sum$ is given by (see Van den Brakel and Renssen (1996c) for a derivation):

$$\hat{\sum} = \mathrm{diag}\left(\frac{\hat{S}_1^2}{n_1}, ..., \frac{\hat{S}_K^2}{n_K}\right), \tag{4}$$

where

$$\hat{S}_k^2 = \frac{1}{(n_k - 1)}\sum_{i\in s_k}\left(\frac{ny_{ik}}{N\pi_i} - \frac{1}{n_k}\sum_{i\in s_k}\frac{ny_{ik}}{N\pi_i}\right)^2, k = 1, 2, ..., K. \tag{5}$$

The expression for the test statistic W can be further simplified to:

$$W = \sum_{k=1}^{K}\frac{n_k}{\hat{S}_k^2}\hat{\bar{Y}}_k^2 - \frac{1}{\sum_{k=1}^{K}\frac{n_k}{\hat{S}_k^2}}\left(\sum_{k=1}^{K}\frac{n_k}{\hat{S}_k^2}\hat{\bar{Y}}_k\right)^2. \tag{6}$$

Note that the $\hat{S}_k^2 / n_k$ are ordinary variance estimators for the sample means, treated as if the sample elements are selected with unequal probabilities (π_i / n) with replacement. These variance estimators depend only on the first order inclusion expectations. No second order inclusion expectations are required. This implies that test statistic (6) is relatively simple to evaluate.

Note that $\hat{S}_k^2$ is an estimate for the population variances for the y_{ik} variables weighted with a factor $n/(N\pi_i)$. If it is reasonable to assume that these weighted population variances of all of the K subsamples are equal, then an efficient estimate is obtained by using the pooled variance estimator:

$$\hat{S}^2 = \frac{1}{n - K}\left(\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(\frac{ny_{ik}}{N\pi_i} - \frac{1}{n_k}\sum_{i=1}^{n_k}\frac{ny_{ik}}{N\pi_i}\right)^2\right). \tag{7}$$

If the pooled variance estimator is used, then test statistic $W/(K - 1)$ corresponds with the F-statistic, whereby the observations y_{ik} are weighted with $n/(N\pi_i)$. In the case of a self-weighted sampling design, $W/(K - 1)$ is equal to the F-statistic, regardless of the second order inclusion expectations of the sampling design.

In the analysis of the experiment, advantage of auxiliary information can be taken by applying the generalized regression estimator instead of the Horvitz-Thompson estimator for the estimation of the parameters from test statistic (2). This increases the precision of the analysis and corrects, at least partially, for the bias due to selective nonresponse. Note that this approach very much resembles the application of covariance analysis from the theory of experimental designs.

If the model assisted approach of Särndal *et al.* (1992) is followed, the target variables for each element in the population are to a certain extent assumed to be an independent realization of a linear regression model. To describe the target variables measured by means of one of the K treatments, K different regression models are defined as follows:

$$y_{ik} = z_i^t b_k + e_{ik}, \; V(y_{ik}) = \sigma_{ik}^2,$$

$$i = 1, 2, ..., N, k = 1, 2, ..., K, \tag{8}$$

where z_i is a vector with values of element i of q auxiliary variables, $\boldsymbol{b}_k$ a vector containing q regression coefficients, e_{ik} the residuals with variances σ_{ik}^2. The generalized regression estimator for $\bar{Y}_k$ based on subsample s_k is

$$\hat{\bar{Y}}_{R_k} = \hat{\bar{Y}}_{\pi_k} + \hat{\boldsymbol{b}}_k^t(\bar{\boldsymbol{Z}} - \hat{\bar{\boldsymbol{Z}}}_{\pi_k}), \; k = 1, 2, ..., K, \tag{9}$$

where $\hat{\boldsymbol{b}}_k$ is the generalized regression estimator of the regression coefficient $\boldsymbol{b}_k$ based on the sampling units in subsample s_k (see Särndal *et al.* (1992), equation 6.4.13), $\bar{\boldsymbol{Z}}$ a vector with the q population means of the auxiliary variables and $\hat{\bar{\boldsymbol{Z}}}_{\pi_k}$ a vector with the Horvitz-Thompson estimators of the population means of the q auxiliary variables based on subsample s_k (with first order inclusion expectation $(n_k / n)\pi_i$. The test statistic (2) based on the generalized regression estimator is

$$W = \sum_{k=1}^{K}\frac{n_k}{\hat{S}_{E_k}^2}\hat{\bar{Y}}_{R_k}^2 - \frac{1}{\sum_{k=1}^{K}\frac{n_k}{\hat{S}_{E_k}^2}}\left(\sum_{k=1}^{K}\frac{n_k}{\hat{S}_{E_k}^2}\hat{\bar{Y}}_{R_k}\right)^2, \tag{10}$$

where

$$\hat{S}_{E_k}^2 = \frac{1}{(n_k - 1)}\sum_{i\in s_k}\left(\frac{n\hat{e}_{ik}}{N\pi_i} - \frac{1}{n_k}\sum_{i\in s_k}\frac{n\hat{e}_{ik}}{N\pi_i}\right)^2, k = 1, 2, ..., K, \tag{11}$$

and $\hat{e}_{ik} = y_{ik} - z_i^t\hat{\boldsymbol{b}}_k$. The derivation of this test statistic resembles the derivation of expression (6) in the case that (2) is based on the Horvitz-Thompson estimator. As an alternative, the g weights (Särndal *et al.* 1992), equation (6.5.10)) can be attached to the residuals in (11).

If it is reasonable to assume that the weighted population variance for the K treatments are equal, then it is more efficient to use the pooled variance estimator. This estimator has the same form as (7) with y_{ik} replaced by respectively $\hat{e}_{ik}$. Instead of defining K separate regression models for each of the K treatments in the experiment, it is also possible to assume that the regression coefficients of the auxiliary variables in each of the K treatments are equal ($\boldsymbol{b}_1 = \boldsymbol{b}_2 = ... = \boldsymbol{b}_K = \boldsymbol{b}$). Then the target variables in the population can be described with one linear regression model. Consequently, the estimates of the regression coefficients $\hat{\boldsymbol{b}}$, based on sample s (with first order inclusion expectation π_i), will be more accurate. Vector $\hat{\boldsymbol{b}}$ can be substituted, in the generalized regression estimator (9).

A special case of the CRD is the analysis of the two sample problem, which aims to test whether there is a specified or unspecified significant difference between two parameters $\bar{Y}_1$ and $\bar{Y}_2$. Hypotheses can be tested by means of the t-test. In Van den Brakel and Renssen (1996b, 1998a) a design based method for the analysis of the two sample problem is proposed by deriving an alternative t-statistic in an equivalent way as is described here for the Wald statistic.

In order to test hypothesis (1), the probability distribution of test statistic (2) has to be known. In the case of simple random sampling without replacement, it can be derived that the limit distribution of test statistic (2) tends to the chi-squared distribution with $K - 1$ degrees of freedom (Van den Brakel and Renssen 1996c). No limit distribution for this test statistic is known for more complex sampling designs, but empirically it can be shown that test statistic (2) has an approximate chi-squared distribution with $K - 1$ degrees of freedom for reasonable large sample sizes in a variety of complex design situations for the two-sample problem.

4. DISCUSSION

Statistical methods from experimental designs and sampling theory can be combined in order to develop efficient methods for design and analysis of experiments embedded in ongoing surveys. Principles of experimentation should be applied in the design and analysis of embedded experiments as much as possible in order to improve the precision of the estimated treatment effects and to avoid that the cause-effect relationship between treatments and observed effects is distorted. Parallels between structures of experimental designs and sampling theory can be exploited in the design of efficient experiments based on these principles in a straightforward manner. The structure of the survey design forms a prior framework for the design of the experiment, *e.g.*, local control by means of randomization within strata, clusters, primary sampling units (PSU's) or interviewers.

To draw inference on finite population parameters of the survey a design based analysis is advocated in this paper. From the objective of the experiment, sensible hypotheses about finite population parameters can be formulated. Efficient test statistics which take into account the probability structure imposed by the chosen sampling design as well as the randomization applied to assign experimental units to the different treatments can be constructed. Such a design based analysis for the embedded completely randomized design and the two sample problem is derived in this paper. Generally, it will be efficient to exercise local control over sampling structures by *e.g.*, randomization within strata, interviewers, clusters or PSU's in two stage sampling designs. Therefore we have also developed a design based analysis for embedded randomized block designs with these sort of sampling structures as (random) block variables (Van den Brakel and Renssen 1998b). This naturally leads to statistical procedures for the design and analysis of embedded experiments which combine the internal validity guaranteed by methods from randomized experimentation with the external validity obtained from the theory of randomized sampling.

REFERENCES

Fienberg, S.E., and Tanur, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review, 55*, 1, 75-96.

Fienberg, S.E., and Tanur, J.M. (1988). From the inside out and the outside in: Combining experimental and sampling structures. *The Canadian Journal of Statistics*, 16, 2, 135-151.

Fienberg, S.E., and Tanur, J.M. (1989). Combining cognitive and statistical approaches to survey design. *Science*, 243, 1017-1022.

Särndal, C.-E., Swenson B., and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Skinner, C.J., Holt D., and Smith, T.M.F. (Eds.) (1989). *Analysis of Complex Surveys*, New York: John Wiley.

Van den Brakel, J.A., and Renssen, R.H. (1996a). Application of experimental designs in survey methodology. *Proceedings of the International Conference on Survey Measurement and Process Quality, American Statistical Association*, 151-156.

Van den Brakel, J.A., and Renssen, R.H. (1996b). The analysis of the two sample problems embedded in complex sampling designs. Research Paper, (BPA no. 1027-96-RSM), Department of Statistical Methods, Statistics Netherlands.

Van den Brakel, J.A., and Renssen, R.H. (1996c). The analysis of completely randomized designs embedded in complex sampling designs. Research Paper, (BPA no. 8090-96-RSM), Department of Statistical Methods, Statistics Netherlands.

Van den Brakel, J.A., and Renssen, R.H. (1998a). Design and analysis of experiments embedded in sample surveys. *Journal of Official Statistics*, in press.

Van den Brakel, J.A., and Renssen, R.H. (1998b). Design and analysis of completely randomized designs and randomized block designs embedded in complex sampling designs. Research paper, in press, Department of Statistical Methods, Statistics Netherlands.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

MODELLING EDUCATIONAL QUALIFICATION FROM THE BRITISH HOUSEHOLD PANEL STUDY

D.M. dos Santos[1] and D.M. Berridge

ABSTRACT

This paper presents a method of analysis for education histories comprising an ordered repeated categorical outcome. We use a generalisation of the continuation ratio model with a random effect. We handle residual heterogeneity by using a normal mixture distribution which allows for stayers, and by using a nonparametric mixture distribution.

KEY WORDS: Ordered repeated categorical outcome; Residual heterogeneity; Continuation ratio mixture model.

1. INTRODUCTION

An individual's education history may be represented by a sequence of levels of qualification, *i.e.*, no qualifications; 0 levels or equivalent; A levels or equivalent; degree or equivalent, that are inherently ordered in some way and that are recorded repeatedly over time. We would like to investigate how a group of explanatory variables influences educational qualification. In order to do this we have to take into account not only ordinality within the levels of educational qualification, but also the usually substantial variation between individuals due to unmeasured and possibly unmeasurable variables (commonly known as residual heterogeneity in the social sciences).

Repeated binary outcomes are just a special case of repeated nominal and ordinal outcomes, when the number of categories is restricted to two. Methods relating repeated binary outcomes to a set of explanatory variables have already been developed. One such approach handles residual heterogeneity by incorporating individual-specific normal error structure into an event-specific binary logistic model (Allison 1987). The residual heterogeneity is eliminated by applying marginal likelihood techniques. An alternative nonparametric approach does not assume a distribution for the random effect.

There are many generalisations of the binary logistic model for a single ordinal outcome, including the continuation ratio model (Fienberg 1977; Fienberg and Mason 1978). In this paper we develop methodology for repeated ordered outcomes by generalising a logistic mixture model for repeated binary outcomes in a similar manner. A major practical advantage of the continuation ratio model is that it can be fitted with ordinary logistic regression programs merely by rearranging the data (Armstrong and Sloan 1989).

In Section 2 we describe the data used to illustrate the approach. Section 3 presents the model specification, while Section 4 discusses the results. The limitations of this methodology are considered in Section 5.

2. DATA

The data employed in this paper have been taken from the British Household Panel Study, BHPS. Respondents were asked to state their highest educational qualification in each of four waves (consecutive years). The levels of qualification, with their corresponding codes, are: 1: no qualification; 2: other qualifications, *e.g.*, technical, trade apprenticeships; 3: O levels or equivalent (up to 16 years); 4: A levels or equivalent (up to 18 years); 5: degree or equivalent (up to 21/22 years).

We would like to relate level of highest educational qualification to the following variables: age (in years) on 1st December, *e.g.*, 01/12/91 (wave 1); sex (0 = male; 1 = female); marital status (married, living as couple, divorced, separated, never married); job status (self employed, in paid employment, unemployed, not employed); number of children in household (from 0 to 5).

The sample data were extracted as follows. The model cannot handle missing values, so only respondents who gave a full interview in each of the four waves were selected initially. Of these respondents, we selected those aged between 16 and 25 years in wave 1, as the age group most likely to see substantial variation in highest educational qualification. Though likely to introduce potential bias into the sample, but to avoid the added complication of multiple respondents per household, we chose the responsible adult in each household in wave 1 and followed that respondent through subsequent waves. After excluding any respondent with missing data, the BHPS sample comprised 1121 respondents.

3. MODEL SPECIFICATION

3.1 Binary Logistic

Assume initially that the repeated outcome is binary rather than ordinal in nature. Consider the binary outcome y_{it} of the t-th wave for respondent i, which equals 0 or 1 for

[1] D.M. dos Santos, Universidade Federal Fluminense, Instituto de Matemática, Rua Mario Santos Braga, s/n-Centro-Niterói-CEP: 24.020-140, Brasil.

all *i* and *t*. There are respondents for whom there will be zero (or very low) probability of a change in level of educational qualification from one wave to another. Such respondents are called 'stayers'. For the moment, it is assumed that there are no stayers. The conventional logistic model can be employed to model such repeated binary outcomes, in which each observation is assumed to be independent. The *t*-th wave associated with the *i*-th respondent contributes the term

$$L_{it}(\beta) = \frac{[\exp(\beta' X_{it})]^{y_{it}}}{1 + \exp(\beta' X_{it})}$$

to the overall likelihood, where X_{it} is a vector of explanatory variables and β is a vector of parameters. A respondent-specific random error term is incorporated into the above model in order to take residual heterogeneity into account. The error term is eliminated by integrating over the likelihood. Assuming a normal mixture distribution, the likelihood integral can be evaluated numerically using Gaussian quadrature. The vector of parameters β, along with their corresponding standard errors, may be estimated using the package SABRE (Software for the Analysis of Binary Recurrent Events) developed by Barry *et al.* 1990. Details on the latest release of SABRE may be found on the World Wide Web at the site: http://www.cas.lancs.ac.uk/software/sabre.html. Repeated binary outcomes may be analysed using other statistical software packages, but SABRE is the only package that has been designed specifically for the modelling of repeated binary outcomes, and that can handle stayers.

3.2 Continuation Ratio

The approach described above can be generalised so that we can investigate whether level of educational qualification depends on age, sex, marital status, job status and number of children in the household. Again it is assumed there are no stayers. The issue of handling stayers when modelling ordinal recurrent events is discussed in Subsection 3.4.

The ordinal outcome in this case comprises five levels: no qualifications, other qualifications, O levels or equivalent, A levels or equivalent, degree or equivalent, coded 1,...,5, respectively. For each respondent, we have a sequence of ordinal outcomes. Suppose that for the *i*-th respondent we have the following sequence of ordinal outcomes:

3 4 5

A natural way of modelling education histories such as the one above would be to consider the original outcome as a series of conditionally independent binary outcomes, each of which may be modelled via binary logistic regression. This partitioning reflects the way in which individuals decide whether or not to continue with their education at each level of schooling. Indeed, an early use of this approach was to assess the effects of parental socio-economic characteristics and family structure on the sequence of school continuation decisions (Mare 1980). Hence, this approach has become known commonly as the continuation ratio model.

Assume respondent *i* reaches level $j(i,t)$ by the *t*-th wave. Then, the probability of this respondent reaching level *j*, given the level is *j* or higher, is

$$\frac{\exp(\theta_j + \beta' x_{it})}{1 + \exp(\theta_j + \beta' x_{it})}; \quad j = 1, \dots, 4$$

The parameter θ_j is the intercept (cutpoint) specific to the *j*-th partition of the original outcome, $j = 1, \dots, 4$. This model is known as the continuation ratio model.

3.3 Logistic-Normal

Denote the outcome corresponding to the *j*-th partition of the *t*-th wave for the *i*-th respondent by y_{ijt}, which equals 1 if *j* equals $j(i,t)$, and equals 0 otherwise, for $j(i,t) = 1, \dots, 4$, and is undefined for values of *j* greater than $j(i,t)$, if $j(i,t) = 1, \dots, 4$. This variable equals 0 for all values of *j*, if $j(i,t) = 5$.

The above sequence of ordinal outcomes can be expressed in terms of y_{ijt}'s as follows:

wave no., *t*	1	2	3
ordinal outcome	3	4	5
y_{i1t}	0	0	0
y_{i2t}	0	0	0
y_{i3t}	1	0	0
y_{i4t}		1	0

The likelihood of respondent *i* attaining level $j(i,t)$ by the *t*-th wave is

$$L_{ij(i,t)t}(\theta, \beta, \alpha; x_{it}, v) =$$

$$\prod_{j=1}^{j(i,t)} \frac{[\exp(\theta_j + \beta' x_{it} + \alpha v)]^{y_{ijt}}}{1 + \exp(\theta_j + \beta' x_{it} + \alpha v)}, \quad j(i,t) = 1, 2, 3, 4$$

$$L_{i5t}(\theta, \beta, \alpha; x_{it}, v) = \prod_{j=1}^{4} \frac{[\exp(\theta_j + \beta' x_{it} + \alpha v)]^{y_{ijt}}}{1 + \exp(\theta_j + \beta' x_{it} + \alpha v)}, j(i,t) = 5$$

The parameter α and variable v are used to handle residual heterogeneity: respondent-specific random error can be incorporated into the above framework to produce a full integrated likelihood of the form

$$L_i(\theta; \beta; x_{it}) = \int \left[\prod_{t=1}^{T_i} L_{ij(i,t)t}(\theta, \beta, 1; x_{it}, \varepsilon) \right] f(\varepsilon)\, d\varepsilon$$

where $f(\varepsilon)$ is the probability density function (mixture distribution) of the error term and T_i is the number of waves for respondent *i*.

As with the binary model, we can use Gaussian quadrature to numerically evaluate the above likelihood integral. Assuming that $f(.)$ takes a normal parametric form, the integrated likelihood becomes

$$L_i(\theta,\beta,\omega;x_{it}) = \sum_{q=1}^{Q}\left[\prod_{t=1}^{T_i} L_{ij,(i,t)t}(\theta,\beta,\omega;x_{it},z_q)\right]P_q \qquad (1)$$

where z_q and P_q, $q = 1,\ldots,Q$, are the Q fixed quadrature locations and probabilities respectively, and the scale parameter ω is the unknown standard deviation of the mixture distribution.

Information on all waves relating to a respondent must be transformed before using likelihood (1). Consider the t-th wave for respondent i, with attained level $j(i,y)$. For this wave, a j-th row of data is generated with the variables:

case: a variable identifying the respondent;
cutp: a variable with value j, representing the j-th partition of the original outcome; j = 1,2,3,4;
out: the variable y_{ijt};
age: age of the respondent on 1st December of the t-th wave;
sex: 1 for females and 0 for males;
mastat: marital status;
jobstat: job status;
nchild: number of children in the household.

By running the whole extended dataset through SABRE, as though modelling repeated binary outcomes, model (1) is actually fitted to the repeated ordered categorical outcomes in the original set of data. The variable *out* is the binary outcome variable. If the variable *cutp* is defined as a factor at four levels and fitted as the only explanatory variable (the null model), then the regression coefficient associated with *cutp(j)* will correspond to θ_j, j = 1,2,3,4. The effects of the explanatory variables may then be investigated.

3.4 Logistic-Normal with Stayers

Assuming that there are stayers, the quadrature points may be supplemented by end points at plus and minus infinity. In the current example, a stayer refers to a respondent who either always has no qualification or always has a degree or equivalent, giving the sequence likelihood:

$$L_i^* = \rho_1\left[\prod_{t=1}^{T_i} y_{i1t}\right] + \rho_5\left[\prod_{t=1}^{T_i} (1 - y_{i4t})\right] + \frac{L_i}{1+\psi_0+\psi_1} \qquad (2)$$

where L_i is given by equation (1),

$$\rho_1 = \frac{\psi_0}{1+\psi_0+\psi_1} \text{ and } \rho_5 = \frac{\psi_1}{1+\psi_0+\psi_1}$$

are estimated proportions of the sample always with no qualifications and always with a degree or equivalent, respectively. SABRE can also be used to fit this model.

3.5 Logistic-Nonparametric

As the normal error distribution is not always robust, it may be prudent to use a nonparametric mixture distribution. The nonparametric maximum likelihood estimate is a discrete distribution on a finite number of mass points (Kiefer and Wolfowitz 1956; Laird 1978; Lindsay 1983). The difficulty of this method lies in determining the location of the mass points. GLIM macros, that use an EM algorithm for maximum likelihood estimation in generalised linear models with random effects, have been written (Aitkin 1996). The algorithm is first derived for Gaussian quadrature assuming a normal mixture distribution, and then modified slightly to handle a nonparametric mixture distribution. The likelihood is algebraically similar to the one given in equation (1) but now the z_q's and p_q's are no longer specified externally and have to be estimated. The scale parameter is absorbed into the nonparametric representation. The number of mass points is also unknown but is increased until the likelihood is maximized.

4. RESULTS

The results in Table 1 demonstrate that ignoring heterogeneity does lead to underestimated (and, in the case of *jobstat*(4), misleading) effects of exogenous explanatory variables and their standard errors. The estimates of ω in the parametric mixture models are highly significantly different from zero, indicating that there is substantial residual heterogeneity, due to the omission of important explanatory variables (exogenous characteristics such as income, and endogenous personal characteristics such as motivation), which has been taken into account by incorporating a random effect into the modelling framework. The proportions of the sample always with no qualifications, and always with a degree or equivalent, are estimated to be 5.5% and 6.4% respectively. These proportions are significantly different from zero, as indicated by the significance of the endpoint parameters.

The extra variation explained by the nonparametric mixture model is due to the additional flexibility in the nonparametric specification of the mixture distribution. The negative estimate for *age* means that the older the respondent, the lower is the conditional probability of having no qualifications; in other words, the higher is the highest educational qualification. The significant negative estimate for *mastat*(5) means that a respondent who has never married proceeds to obtain higher qualifications at a faster rate than a married respondent. The significant positive estimate for *jobstat*(4) in the nonparametric mixture model indicates that the conditional probability of having no qualifications is higher for respondents not employed than those who are self-employed. The positive estimate of *nchild* implies that the more children a respondent has, the higher is the conditional probability of having no qualifications; in other words, the lower is the highest educational qualification. The negative estimate for *sex*, which only becomes significant under the nonparametric mixture model, indicates that female respondents proceed to higher qualifications faster than male respondents.

Table 1
Results of fitted models

PARAMETER	continuation ratio	logistic-normal (without stayers)	logistic-normal (with stayers)	logistic-nonparametric
	estimate (s.e.)	estimate (s.e.)	estimate (s.e.)	estimate (s.e.)
cupt(1)	0.352 (0.284)	-2.533 (0.540)	-2.678 (0.549)	16.32 (4.587)
cutp(2)	1.004 (0.283)	1.417 (0.540)	1.372 (0.552)	20.80 (4.594)
cutp(3)	2.546 (0.284)	6.654 (0.564)	6.324 (0.577)	26.18 (4.601)
cutp(4)	3.308 (0.289)	10.390 (0.587)	10.198 (0.594)	30.19 (4.608)
age	-0.127 (0.009)	-0.266 (0.019)	-0.255 (0.020)	-0269 (0.017)
mastat(2)	-0.161 (0.078)	-0.260 (0.152)	-0.269 (0.152)	-0279 (0.140)
mastat(3)	0.259 (0.204)	-0.366 (0.340)	-0.247 (0.370)	-0.025 (0.351)
mastat(4)	1.103 (0.216)	-0.091 (0.381)	-0.077 (0.404)	0.002 (0.408)
mastat(5)	-0.350 (0.070)	-0.609 (0.143)	-0.521 (0.149)	-0.409 (0.124)
jobstat(2)	-0.210 (0.134)	-0.187 (0.274)	-0.132 (0.301)	-0.142 (0.244)
jobstat(3)	0.215 (0.150)	-0.480 (0.305)	-0.486 (0.330)	-0.498 (0.273)
jobstat(4)	·0.223 (0.138)	0.403 (0.289)	0.470 (0.132)	0.708 (0.252)
nchild	0.423 (0.029)	0.174 (0.059)	0.137 (0.061)	0.092 (0.052)
sex				-0.164 (0.049)
ω		4.261 (0.094)	3.977 (0.101)	
end point 1			0.062 (0.009)	
end point 5			0.073 (0.016)	
ρ_1			0.055	
ρ_5			0.064	
-2log-lik.	12739.426	8184.044	7985.272	7890.800
-2log-lik. (null model)	13295.290	8507.565	8294.088	8259.106

s.c. = standard error, -2log-lik. = -2 × log-likelihood

5. DISCUSSION

It is implausible that a univariate error distribution as the normal, employed in this paper, can represent the effects of residual heterogeneity in an effective manner, with different processes likely to govern moves out of different states.

We have assumed here that the mixture distribution of the random effect is identical over all partitions of an ordinal outcome. This assumption is likely to be unrealistic since the number of outcomes in any one partition decreases with successive partitions. The nonparametric approach can handle partition-specific mixture distributions by way of an interaction between the factor *cutp* and the factor used to represent the mass points in the model.

REFERENCES

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, 6, 251-262.

Allison, P.D. (1987). Introducing a disturbance into logit and probit regression models. *Sociological Methods and Research*, 15, 355-374.

Armstrong, B.G., and Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, 129, 191-204.

Barry, J., Francis, B.J., and Davies, R.B. (1990). SABRE: *Software for the analysis of binary recurrent events. A guide for users.* CAS Publications, Centre for Applied Statistics, Lancaster University.

Fienberg, S.E. (1977). *The analysis of cross-classified categorical data.* Cambridge, Mass: MIT Press.

Fienberg, S.E., and Mason, W.M. (1978). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. Sociological Methodology 1979, (Ed. K.E. Schuessler), San Francisco, Jossey-Bass, 1-67.

Kiefer, J., and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Annals of Mathematical Statistics*, 27, 887-906.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73, 805-811.

Lindsay, B.G. (1983). The geometry of mixture likelihoods, part I: a general theory. *The Annals of Statistics*, 11, 86-94.

Mare, R.D. (1980). Social background and school continuation decisions. *Journal of the American Statistical Association*, 75, 295-305.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, 42, 109-142.

Proceedings of Statistics Canada Symposium 97
New Directions in Surveys and Censuses
November 1997

COMPARISON OF INSTITUTIONALIZED AND NON-INSTITUTIONALIZED PERSONS WITH LIMITATIONS IN CANADA

C. Houle,[1] J.-M. Berthelot and R. Roberge

ABSTRACT

Currently only 1% of the Canadian population, all ages combined, reside in health care institutions. This proportion rises to 5% for individuals aged 65 and over, and it soars to 18% for persons aged 80 and over. This means that with an ageing population, institutionalization will in future exert increasing financial pressure on the health care system. With data from the National Population Health Survey, individuals aged 65 and over can be classified according to their level of activity limitation. This classification opens the door to research into the factors that explain differences in type of accommodation (institution vs private household) for senior citizens with an equivalent level of disability. The results of logistic regressions indicate that long-term illness, living alone and diversity of income sources are important explanatory variables. Policy makers can draw on these findings and use them to develop appropriate social and health policies.

KEY WORDS: Institutionalization; Activity limitation; Health; Health index; National survey.

1. INTRODUCTION

Currently only 1% of the Canadian population, all ages combined, reside in health care institutions. This proportion rises to 5% for individuals ages 65 and over, and it soars to 18% for persons aged 80 and over (Tully 1995). By using population projections for Canada (Statistics Canada 1994) and institutionalization rates by sex (Tully 1995), it can be projected that the number of beds in institutions could rise from 185,600 in 1994-95 to more than 565,000 in 2031. The prospect of tripling the number of beds available in institutions has major consequences, not only in terms of the financial pressure on the health care system but also in terms of the organization of care. Human resources training, construction of facilities and the possible transfer of services to outpatient care leave little room for improvisation.

In order to inform policy makers who must establish appropriate health policies, it is important to define a conceptual framework and understand the structural differences that explain why some senior citizens are in institutions while others remain in private households. On this subject, we know that even though almost all senior citizens living in health care institutions experience some activity limitation, there is also a sizable proportion of senior citizens living in private households who also experience an activity limitation (Dunn 1990). The objective in classifying these persons by level of disability is to identify the factors that explain the differences in type of accommodation for senior citizens, taking level of disability into account.

2. DATA SOURCE

The data analysed are drawn from the 1994-1995 National Population Health Survey (NPHS) (Tambay 1995). That survey, which lends itself to both longitudinal and cross-sectional analysis, gathered information on 26,400 private households in Canada's provinces. The survey also has an institutional component, consisting of 2,287 residents of 230 health care institutions. Our analysis focuses solely on persons aged 65 and over. There are 1,845 senior citizens in the institutional component and 5,302 in the private household component.

3. METHODS

3.1 Definitions of Disability Levels

Two types of questions were used to define disability levels. Question 46 of the questionnaire for institutional residents was worded as follows: "Because of a long-term physical or mental condition or health problem, are you (is ...) limited in the kind or amount of activity you (he/she) can do?" This question evaluated whether the person was reporting at least one *limitation*. Questions 1a, 1b, 1c, 1d and 2 of the questionnaire for private households separately evaluated limitations at home, at work, in recreational activities and limitations due to a long-term health problem. The second type of questions evaluated *dependency*. Question 47 of the questionnaire for institutional residents and Question 6 for private households was worded as follows: "Because of any long-term condition or health problem, do you (does ...) need the help of another person in (a) personal care? (b) moving about inside the residence?"

The concepts of activity limitation and dependency were used to define three levels of disability: severe, moderate and no disability. Table 1 summarizes the definitions and describes the sample sizes. From it, two major observations may be made: (a) the majority of persons living in institutions have a severe disability, and (b) it is estimated that there are significantly more individuals with a severe disability living in private households than in institutions.

[1] Christian Houle, Statistics Canada, R.H. Coats Building, 24th floor, Ottawa, Ontario, Canada, K1A 0T6; e-mail: houlchr@statcan.ca.

Table 1
Definitions and Sample Sizes

Institutions	Private households
Severe disability • Activity limitation (Q46) *and* dependency Q47 ab • n = 1,241 • N_{est} = 126,098	Severe disability • Activity limitation (Q1 abcd or Q2) *and* dependency (Q6) • n = 291 • N_{est} = 166,771
Moderate disability • Activity limitation but *no* dependency • n = 354 • N_{est}= 35,244	Moderate disability • Activity limitation but *no* dependency • *No* activity limitation but dependency • n = 1,907 • N_{est} = 1,123,112
No disability • *No* activity limitation and *no* dependency • n = 250 • N_{est} = 24,788	No disability • *No* activity limitation and *no* dependency • n = 3,104 • N_{est} = 1,954,606

3.2 What is the Health Utility Index?

The Health Utility Index (HUI) is a generic health status index that is able to synthesize both quantitative and qualitative aspects of health. The HUI provides a quantitative measure of a population's overall health status, as opposed to focussing on narrowly defined health indicators, risk factors or diseases. Such a measure is an analytical tool which can be used for health policy evaluation and for the assessment of health care interventions at a population and a clinical level.

The first step in the construction of a health utility index is specification of a set of health status attributes. These are chosen to reflect a wide range of functional capacities and to capture what most people consider the most serious health-related problems they might encounter. The method developed at McMaster University's Centre for Health Economics and Policy Analysis (CHEPA) describes an individual's overall functional health, based on eight attributes: vision, hearing, speech, mobility (ability to get around), dexterity (use of hands and fingers), cognition (memory and thinking), emotion (feelings), and pain and discomfort. An individual's health status is the vector of the observed levels of functional ability for each attribute (Figure 1). For example, "vision" ranges from perfect vision to blindness, while "pain" ranges from no pain to completely disabling pain.

The HUI maps any one of the vectors of eight health attribute levels into a summary health value with the value 0 representing death and the value 1 representing complete health. For instance, an individual who is near-sighted, yet fully healthy on the other seven attributes, receives a score of 0.95 on this scale.

The HUI value embodies the views of society concerning health status. Societal preferences are defined as an average of the preferences of individuals, insofar as they form a representative sample of the population. Preferences are evaluated using the Standard Gamble technique (Feeny 1995, Torrance 1995), which is based on the axioms of consumer utility theory developed by von Neumann and Morgenstern (1947). They are evaluated on the basis of the Childhood Cancer Study conducted by CHEPA at McMaster University (Feeny 1991).

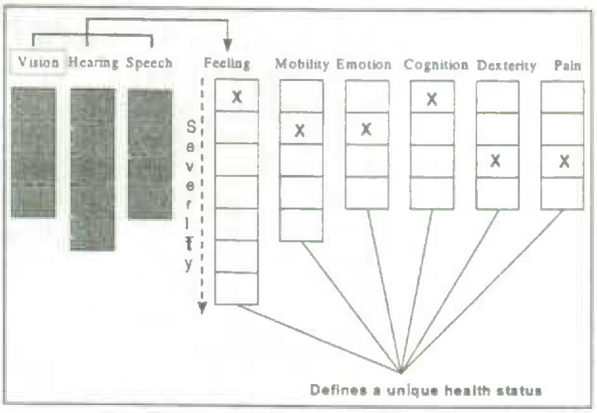


Figure 1. Health Utility Index

3.3 Logistic Regressions

After the survey weight was normalized so as to obtain a mean of 1, the modelling of four logistic regressions was carried out using SAS 6.11. The first regression (I) consists in comparing all individuals in institutions to all individuals living in private households. The second regression (II) compares persons with one severe disability only, according to their type of accommodation. The third (III) does the same thing for persons with a moderate disability. The final regression compares persons living in an institution and reporting no disability to those living in a private household and reporting a severe or moderate disability. In all the regressions, the main interaction factors were tested but none proved to be significant.

4. RESULTS

As Table 2 shows, socio-demographic factors (average age, sex, marital status) exhibit similar patterns according to type of accommodation (institutional or private household), regardless of the level of disability. On the other hand, the general health level and the HUI are comparable by level of disability, regardless of the type of accommodation, although the HUI is in each case slightly higher for persons living in private households than for residents of institutions.

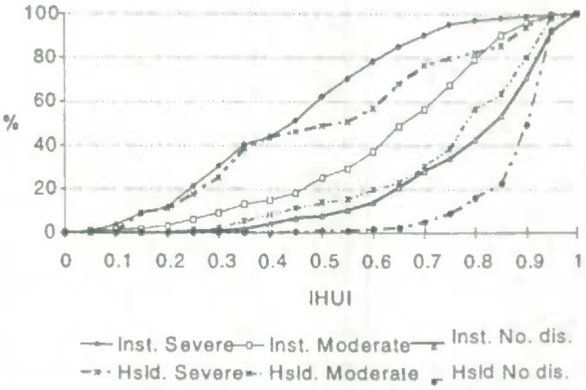
From an analysis of the cumulative HUI curves in Figure 2, two conclusions may be drawn. First, for each accommodation type, the curves are in ascending order of health status for the population that they represent, *i.e.*, the worst states of health (severe disabilities) are furthest to the left, while the best states of health are on the far right. Second, the curves for a given level of disability are relatively similar, regardless of accommodation type. The similarity is especially obvious for severe disability.

Table 2
Demographic Profile and Health Indicators

	Institutions	Private households
Severe	• Average age = 83.7 • Sex: Female = 74% Male = 26% • Marital status: Widowed = 66% Married = 17% • General health • Excellent = 2% Poor = 26% • Average Health Utility Index = 0.43	• Average age = 78.6 • Sex: Female = 61% Male = 39% • Marital status: Widowed = 39% Married = 51% • General health • Excellent = 1% Poor = 26% • Average Health Utility Index = 0.51
Moderate	• Average age = 83.5 • Sex: Female = 71% Male = 29% • Marital status: Widowed = 64% Married = 15% • General health • Excellent = 6% Poor = 11% • Average Health Utility Index = 0.63	• Average age = 73.8 • Sex: Female = 57% Male = 43% • Marital status: Widowed = 31% Married = 59% • General health • Excellent = 4% Poor = 12% • Average Health Utility Index = 0.75
No Dis.	• Average age = 82.0 • Sex: Female = 71% Male = 29% • Marital status: Widowed = 67% Married = 10% • General health • Excellent = 10% Poor = 7% • Average Health Utility Index = 0.79	• Average age = 72.6 • Sex: Female = 56% Male = 44% • Marital status: Widowed = 29% Married = 60% • General health • Excellent = 19% Poor = 1% • Average Health Utility Index = 0.89

Can the HUI help us identify factors that explain a person's presence in an institution for a given disability level? Breaking down the index according to the main attributes tells us about the nature of the disabilities. Figure 3 shows that for the severe disability level, the "pain" component is dominant for individuals living in a private dwelling, whereas the "cognition" and "emotion" components explain the low health level of persons living in institutions. This observation is consistent with the fact that diseases involving considerable pain (such as arthritis) do not necessarily result in institutionalization, whereas degenerative diseases such as Alzheimer's almost always lead to institutionalization, owing to the need to receive care constantly.

Lastly, Table 3 summarizes the results obtained by modelling the four logistic regressions described above. Based on a significance level of 0.01, four major observations are in order. First, the variable "Alzheimer's disease and other dementia" is clearly the one that best explains overall a person's presence in an institution. The order of magnitude of the odds ratio is such that the causal link is beyond doubt. Each logistic regression identifies arthritis as a "protection" factor. It is surprising to see how consistent these findings are with the conclusions drawn regarding the pain and cognitive components of the breakdown of the HUI. Second, we observe that the sex variable is not significant for any regression. Third, when the general regression (I) is compared with the particular regression (IV), it is clear that for institutional residents without disabilities, Alzheimer's disease becomes a secondary (but nevertheless important) explanation, since living alone (approximated by marital status) exhibits the greatest odds ratio. Fourth, income diversification significantly reduces the odds of being in an institution.



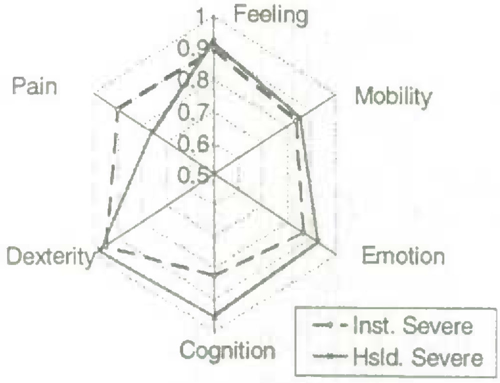
(1) Adjusted for age and sex

Figure 2. Cumulative HUI Curves

Table 3
Results of logistic regressions

Variable	OR Reg. I	OR Reg. II	OR Reg. III	OR Reg. IV
Age	1.12	1.05	1.15	1.11
Widowed	8.66	7.29	9.38	13.8
Other unmarried	4.56	4.08	4.83	7.89
Limitation	NS	N/A	N/A	N/A
Dependency: Personal care	9.37	N/A	N/A	N/A
Dependency: Getting around	1.81	N/A	N/A	N/A
Incontinence	5.84	6.12	5.69	2.18
Arthritis	0.73	0.54	0.56	0.39
Alzheimer's disease and other dementia	17	8.14	18.9	6.43
High blood pressure	0.63	NS	NS	NS
Effects of stroke	2.89	1.93	3.26	NS
Heart disease	NS	NS	NS	NS
Diabetes	NS	NS	NS	NS
Cataracts	NS	NS	NS	NS
Digestive conditions (ulcers)	2.19	2.94	NS	NS
Bronchitis or emphysema	2.27	NS	NS	2.04
Retirement pension income	0.25	0.3	0.25	0.24
Div. or interest income	0.56	NS	0.58	0.38
Other income*	0.15	0.19	0.16	0.09
Sex	NS	NS	NS	NS

* See "other income" category in questionnaires for detailed list.



(1) Adjusted for age and sex

Figure 3. Main HUI Attributes

5. DISCUSSION

In view of current social and economic conditions and the major demographic changes that will occur in the short and medium term, it is time to think about health care policies to apply to the elderly over the next thirty years. The analysis shows that while long-term conditions are the primary cause of institutionalization, there are cofactors that can either help to prevent institutionalization or be useful to consider in planning for it. Diversification of income sources seems to allow individuals to avoid some degree of limitation by purchasing services, and it may be considered a protection factor. Social policies promoting the diversification of income sources may therefore bear fruit.

In addition, for some individuals who appear to have been living on their own, the fact that they did so clearly explains their institutionalization. In particular, senior citizens who report no limitation appear to undergo institutionalization as a direct result of living alone. It is possible that no social or health policy can prevent isolation. If this is true, it is important to assess the legacy of the social and medical changes of the past thirty years. The divorce rate of the 1970s to 1990s suggests that there will be an increase in persons living alone. On the other hand, the opposite effect should result from the shrinking of the gap in life expectancy between men and women, since couples will live together longer. It will also be necessary to look for better indicators that a person is living alone than marital status, since the increase in the number of common law unions will weaken the existing association between this indicator and the concept of individuals living alone. Any model leading to projections regarding the number of institutional beds needed should take these observations into account.

ACKNOWLEDGEMENTS

The authors wish to thank Rolande Bélanger for her technical support, which was effective and highly appreciated. We are also grateful to all members of the NPHS who participated, either directly or indirectly, in this project.

REFERENCES

Dunn, P. (1990). *Barriers confronting seniors with disabilities.* Statistics Canada, Cat. No. 1, 82-615.

Feeny, D., Barr, R.D., Furlong, W. *et al.* (1991). Quality of life of the treatment process in pediatric oncology: An approach to measurement. *Effect of Cancer on Quality of Life. Boca Raton: CRC Press*, 73-88.

Feeny, D., Furlong, W., Boyle, M., *et al.* (1995). Multi-attribute health status classification systems: Health utilities index. *Pharmacoeconomics*, 7(6): 490-502.

Statistics Canada (1994). *Population Projections for Canada, Provinces and Territories 1993-2016.* Cat. No. 91-520.

Tambay, J.L., and Catlin, G. (1995). Sample Design of the National Population Health Survey. *Health Reports* 7(1): 33-42, Statistics Canada, Cat. No. 82-003.

Torrance, G.W., Furlong, W., Feeny, D. *et al.* (1995). Multi-attribute preference functions: Health utilities index. *Pharmacoeconomics*, 7(6): 503-520.

Tully, P., and Mohl, C. (1995). Older Residents of Health Care Institutions. *Health Reports* 7(3): 27-30, Statistics Canada, Cat. No. 82-003.

von Neumann, J., and Morenstern, O. (1947). Theory of games and economic behavior. *Princeton: Princeton University Press.*

CLOSING REMARKS

CLOSING REMARKS

M.P. Singh[1]

Now that we have come to the end of three hectic days of sessions, I have the pleasant responsibility of making a few closing remarks and recognizing the efforts of those who have contributed to the success of this event. And what a grand success it has been.

As Gordon Brackstone noted in his Opening Remarks, this was our fourteenth International Methodology Symposium. In the past, the symposium topics have been more focused on selected issues such as small area estimation, longitudinal surveys, the treatment of nonresponse, *etc*.

The objective of this year's symposium on 'New Directions in Surveys and Censuses' was to bring together specialists engaged in methodological innovations in all aspects of surveys from different parts of the world. As Dr. Fellegi mentioned during his earlier presentation, the topics were chosen to reflect the applied nature of work at Statistics Canada.

This has been our largest symposium. Over 500 participants from a number of countries contributed to its success. There were 22 invited and contributed paper sessions, with 75 papers presented. I know we all had our choices of papers, topics and sessions that we liked most and I am not going to comment on them.

However, when I was asked to make a few remarks, I thought since a large number of papers on diverse topics are being presented, maybe by looking at the titles, abstracts and authorship, I may be able to say something about the goal that the committee had set for the conference: How did it turn out finally? Is there any indication of a specific direction of developments in various parts of the world based on the papers presented at the conference?

In order to do that I grouped the papers under three broad headings on which the committee had focussed the conference, namely,

(i) Program design,

(ii) Data Collection and

(iii) Analysis and Dissemination.

Of the total of 75 papers, there are 30 papers on program design, 20 on data collection and 25 on analysis and dissemination – quite an even spread of papers. After that I cross classified these papers by countries, in 3 groups, namely – European countries (which included a few other countries as well grouped into one), Canada and the United States. Surprisingly, again, a very good distribution of papers emerged. There are 25 papers from Europe, 27 from Canada and 23 from the United States.

I am sure you will agree when I say the committee did an admirable job. I don't know whether it was by chance or by good planning, I guess it was a bit of both, to get such an excellent representation by topic and by countries.

When I looked further at the numbers withing the 3x3 cell (topics by countries), a certain direction of developments as well did emerge. Over half the papers from Europe were on Program Design related issues whereas papers on data analysis and dissemination issues dominated in the case of Canada. Perhaps this is an indication of the emergence of a number of new surveys and integration initiatives in Europe and the recent emphasis on data analysis and longitudinal surveys in Canada. Further, as expected, there was an even distribution on all the 3 topics in the case of United States.

I would like to turn to my more important task and would like to thank all the speakers, discussants and chairs for the time and work that they put into sharing their experiences. First, as they say, charity beings at home. Thanks to the presenters and chairs from Statistics Canada, especially to those from subject matter divisions, who participated in large numbers at this conference.

For those from outside Statistics Canada, I would like to say that we here at Statistics Canada cherish your good will and the honour that you have accorded us by responding so generously to our invitations. Thank you very much.

Then my very sincere thanks go to all those involved in providing various facilities in an admirable manner such as the staff of the Palais des Congrès, the City of Hull for letting us use the very pleasant AGORA for our wine and cheese reception, the staff handling the audio-visual facilities, and lastly but not least, the people doing the simultaneous interpretations, a very difficult job for technical presentations at an international gathering with all kinds of accents.

Special thanks go to the more than 30 volunteers whose work is essential to the success of a conference. They worked hard days and evenings for the last few weeks. But I won't risk mentioning names because I'm certain to miss a few – and I have to get back to the office – so thanks to all the volunteers.

Lastly, and most importantly, thanks go to the Organizing Committee which started working about 20 month ago. It includes Johanne Denis, Dave Dolson, Marc Hamel, Diane Stukel and Kathryn Williams and has brought us a memorable conference under the able leadership of

[1] M.P. Singh, Director, Household Survey Methods Division, 16-O, R.H. Coats Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

Jack Gambino. Let's give them a big hand for a job so well done.

I have now to make some announcements. Next year, we will have a symposium and workshop on Longitudinal Analysis for Complex Surveys, not in the fall this time, but on May 19-22, 1998 at Statistics Canada, organized under the chairmanship of Michael Hidiroglou and Sylvie Michaud. Lastly I can't miss the temptation of putting a bit of a commercial and encourage you not only to subscribe to Statistics Canada's *Survey Methodology* journal and to learn all the new developments taking place around the world, but also I would like to invite you to send articles from this conference and others for possible publication.

This brings the conference to a close – I thank all the participants, and to those from out of town, I wish you a safe journey home.