c3

SURVEY METHODOLOGY

1975-2000

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 1999

•

VOLUME 25

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1999 • VOLUME 25 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

February 2000

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is $47 per year in Canada and US $47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada 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. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 25, Number 2, December 1999

CONTENTS

In This Issue

This December 1999 issue and upcoming June 2000 issue of the Journal contain papers from some prominent statisticians invited to participate in the celebration of 25 successful years of existence of *Survey Methodology*. As an introduction to this special issue, Richard Platek, the founding Chairman, who remained at the helm of the Journal until 1987, has provided an excellent overview of the gradual evolution of the Journal from a modest divisional to a respected departmental publication and then to an international publication of repute.

I would like to take this opportunity to acknowledge a few important events which helped shape the Journal to the status it currently enjoys internationally. Firstly, during the first 10 years of the Journal, which may be called its formative years, important articles by some of the senior management of the Bureau, such as G.J. Brackstone, I.P. Fellegi, P.G. Kirkham, L.E. Rowebottom, J. Spear, M.B. Wilk and D.A. Worton, as well as by some well known survey statisticians, such as J.G. Bethlehem, R.E. Fay, S.E. Fienberg, W.A. Fuller, L. Kish, G. Nathan , J.N.K. Rao and C.-E. Särndal helped to set the solid foundation of the Journal and defined it's role as a forum for publication of innovative articles relevant to a statistical agency.

Secondly, the support and recognition of *Survey Methodology* by the ASA Section on Survey Research Methods, and in particular by some of its past chairs B. Bailar, G. Kalton, F.J. Scheuren and D. Binder, helped to popularize the Journal more widely among survey methods researchers as well as practitioners.

Lastly and perhaps most importantly, a large part of the success the Journal has enjoyed over the years may be attributed to the excellence and dedication of the Editorial Board members and the strong commitment of the referees. The size and composition of the Board has changed significantly over the years. The current membership as usual is given on an earlier page and a complete list of past Associate and Assistant Editors is provided at the end of this issue. There are however a few members, such as D.R. Bellhouse, J.N.K. Rao, and G. Kalton, who joined the Board in 1984, the year when the scope of the Journal was greatly enlarged, and are still strongly committed to its cause.

I now turn to the individual papers in this special issue.

Fellegi considers the challenges facing government statistical agencies and strategies to prepare for these challenges. He first describes the environment of changing information needs and the social, economic and technological developments driving this change. He goes on to describe both internal and external elements of a strategy to meet these evolving needs. Internally, a flexible capacity for survey taking and information gathering must be developed. Externally, contacts must be developed to ensure continuing relevance of statistical programs while maintaining non-political objectivity.

Kish describes the challenges and opportunities of combining data from surveys of different populations. Examples include multinational surveys where the data from surveys of several countries are combined for comparison and analysis, as well as cumulated periodic surveys of the "same" population. He also compares and contrasts the combining of surveys with the combining of experiments.

Brackstone discusses issues of quality in the products of a national statistical agency. He identifies and discusses six different dimensions of data quality: relevance, accuracy, timeliness, accessibility, interpretability and coherence. He then describes the components of a quality management system.

In his paper Scheuren considers the possible uses of administrative records to enhance and improve population censuses. After reviewing previous uses of administrative records in an international context, he puts forward several proposals for research and development towards increased use of administrative records in the American statistical system.

Godambe and Thompson consider the problem of confidence intervals in survey sampling. They first review the use of estimating functions to obtain model robust pivotal quantities and associated confidence intervals, and then discuss the adaptation of this approach to the survey sampling context. Details are worked out for some more specific types of models, and an empirical comparison of this approach with more conventional methods is presented.

J.N.K. Rao gives an overview of the methods and models used for small area estimation. This is an update of his previous overview (Ghosh and Rao, 1994, *Statistical Science*). He first presents a general discussion of small area models, making a distinction between areal level models and unit level models. He then describes the development in the three main approaches for inference based on these models: EBLUP, EB and HB, and gives several examples of recent applications. Finally, he presents an interesting discussion identifying the gaps and areas that require further research.

Sirken and Shimizu derive a Horvitz-Thompson estimator for population based establishment sample surveys (PBESs). A PBES is a survey of establishments where the sampling frame consists of establishments with which a preliminary sample of households or individuals has had some contact.

Deville shows how to use simple tools to calculate the variance of a complex estimator using a linearization technique. The process is that of a software used at INSEE for estimation of the variance of a complex estimator. It gives a way of computing the variance of a total estimated by the simple expansion estimator. In the case of a complex statistic, the process uses a derived variable that reduces the computations to those of the simple expansion estimator. Multiple examples are given to illustrate the process.

Brewer proposes a method of weight calibration in survey sampling, called cosmetic calibration, which yields cosmetic estimators of totals, *i.e.* estimators that can be interpreted as both design-based and prediction based. He also discusses variance estimation and shows how the problem of negative weights can be easily and naturally handled using cosmetic calibration. Finally he compares the properties of the weights and the resulting estimators to some alternative approaches using some Australian farm data.

In the final paper of this special issue, Estevao and Särndal consider two types of design-based estimators used for domain estimation. The first, a linear prediction estimator, is built on the principle of model fitting, requires known auxiliary information at the domain level, and results in weights that depend on the domain to be estimated. The second, a uni-weight estimator, has weights which are independent of the domain being estimated and has the clear advantage that it does not require the calculation of different weight systems for each different domain of interest. These estimators are compared and situations under which one is preferred over the other are identified.

I am pleased to add that with this 25th Anniversary issue, we are making Survey Methodology *available to you in electronic format. It is easy to access the Journal on our Web site by keying in the following URL address: www.statcan.ca/english/e-pub.*

Once you've checked out the electronic version we'd appreciate you completing the brief on-line survey you'll find at the same location. This prototype is being offered as a test to find out what your future format preferences for the Journal might be.

I assure you that any change in delivery format will not affect the high quality you expect and receive from Survey Methodology.

M.P. Singh

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 109-111
Statistics Canada

Survey Methodology – The First 25 Years

RICHARD PLATEK[1]

This year the *Survey Methodology* journal, published by Statistics Canada, celebrates the silver anniversary of its remarkably successful existence that began in 1975. During these 25 years, the journal has developed into a dynamic and innovative leader in survey methodology. The future promises the same.

The first issue of the journal appeared in June 1975. Although the publication of the journal may have surprised many, the period 1960-75 provided a ripe background for it. In Statistics Canada methodological research flourished. Important and challenging ideas were developed worthy of presentation at international conferences. In this period of productive activities, *Quarterly Bulletin* and *Memoranda*, the forerunners of the journal, were established. These publications highlighted more significant developments, and were intended as a tool in staff training. Their circulation had been mostly internal, however, they soon found their way outside to various statistical organizations, universities and private research centres. The breadth of subject matter covered by the two series was enormous. They provided a springboard for a number of important papers on topics such as small area estimation, record linkage, edit and imputation, variations in response and others. As research and development work kept increasing, so too did a need and desire among statisticians at Statistics Canada to publish formal papers that could be subjected to a refereeing process. Thus, in 1975, the *Survey Methodology* journal was quietly but not prematurely born. At the beginning of its existence, the journal was for all intents and purposes an in-house publication guided and nurtured by one division.

Notwithstanding the journal's modest beginning, its first Editorial Board, consisting of Richard Platek (Chairman), M.P. Singh (Editor), and Paul Timmons, were not modest in their ambitions, as expressed in the editorial policy:

> "The objective of the Survey Methodology journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology journal will publish articles dealing with all phases of methodological development in surveys, such as design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and applications, statistical analysis, interpretation, evaluation and inter-relationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed, however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of the Department."
> (Statistics Canada 1975)

The foregoing makes it abundantly clear that the intent was to create a professional journal with a very specific and unique focus. Up to this point, there had not existed a forum for the illumination of general methodological issues arising during the course of putting a survey into the field. Most statistical journals were accepting articles dealing primarily with the mathematical aspects of sample design. The journal filled a long existing vacuum in the field of survey methodology. It provided an immediate benefit and challenge to survey methodologists. By publishing in the journal they had the opportunity to disseminate their work and ideas to a wider range of survey practitioners and theoreticians. In a number of countries, the journal, almost from the beginning, provided a base for teaching and training new statisticians. As S.S. Zarkovich, a prominent statistician from Yugoslavia put it:

> "In this country, the situation is as follows. Survey Methodology was a highly regarded periodical by the young generation of statisticians in this country. It was a subject of conversation. A good part of production of our young generation was based on the ideas expressed in Survey Methodology." (Zarkovich 1985).

Initially, for a number of years, the journal relied exclusively on papers written by the staff in Statistics Canada. It was distributed, free of charge, within and outside Statistics Canada. The external recipients were: federal and provincial departments, university libraries, survey research centres, and statistical organizations abroad. This gracious policy had far reaching effects. The journal received a great deal of support from statisticians and statistical organizations in many countries. Statisticians began to enquire about submitting their papers and statistical organizations expressed interest in subscribing to the journal. In a relatively

[1] Richard Platek, International Consultant on Statistical Surveys, formerly at Statistics Canada, Ottawa, Ontario, Canada.

few years of its existence the journal won and secured for itself professional recognition and international stature as a technical journal with a unique focus on survey methodology. References to and abstracts from the papers published in the journal now appear in various statistical journals and publications. A brief article on the journal was featured in Encyclopedia of Statistical Sciences (Singh 1988).

Over the first 10 years the journal was steadily and successfully evolving but the process was gradual. The year 1984 saw the result of many important decisions that had the potential to change the character of the journal significantly. First of all the journal became an official Statistics Canada publication and with this came several concomitant developments. Commensurate with Statistics Canada's policy for its publications, the journal acquired a price tag, and officially became a bilingual publication (Wilk 1982). The general appearance, printing, and format were improved. The production process became smoother and more efficient to ensure timeliness and quality of the final product. With respect to editorial policy and the scope of the journal, it was realized that a broader base in its content, its contributors and its readership would enhance its value and effectiveness. On the other hand, there was some concern that its main objectives, as expressed in the initial Editorial Policy, not be lost in the process. Finally, a decision was made to broaden the journal's scope, expanding the Editorial Board to new areas, accepting and inviting papers from outside. The international community responded to the invitation by submitting many papers to the journal. This trend continues to this day.

Another important decision was the establishment of a separate Management Board. The Management Board, with the help of a Production Manager would coordinate various phases of the journal's production to keep pace with its professional responsibilities. The journal's pricing, relationship with other journals, and production issues are frequently on the agenda of the Management Board.

Concurrently with all the new decisions, discussions were held and distribution agreements were reached with the International Association of Survey Statisticians, the Statistical Society of Canada, and the American Statistical Association, and more recently with the American Association for Public Opinion Research. The members of these associations were given the opportunity to subscribe to the journal at various special rates.

All of these developments resulted in a different perception of the journal, both internally and externally. Externally, the journal entered the international arena. Internally, important changes took place in its Editorial Policy, which emphasized that "The Survey Methodology journal will publish articles dealing with various aspects of statistical development relevant to a statistical agency" (Statistics Canada 1984).

While keeping abreast of new requirements and needs, the journal continues to extend its scope to cover a full range of methodological questions arising in surveys. In its steady growth, the journal introduced, in 1988, a preface called "In this Issue" (Statistics Canada 1988). Furthermore, from time to time, the journal has been carrying special sections dedicated to topics of particular interest. Examples of topics that were covered in such special sections include census coverage errors, data analysis, establishment survey methods, and longitudinal surveys and analysis. Notices of conferences and seminars on surveys also appear in the journal.

To reflect the increasing number and depth of topics published in the journal and to keep pace with its professional commitments, the journal's Editorial Board was enlarged and another new Management Board was established.

The very essence of any journal is its sensitivity to changing times and expectations. The management of the journal has been, therefore, conducting market research studies in order to evaluate customers' reaction to the journal. Although the outcome of some recent investigations show that some readers feel that their needs for practical applications are not fully met, most regard the journal as a high quality journal and are satisfied with its content.

In recent years the journal has become very popular among academics, with many seeking to publish their articles in it. A view has been expressed that, if this is not controlled, the journal may lose its general and educational appeal to some readers in many countries. While it is proper and healthy to recognize and publish theoretical papers, it is equally important to ensure that this will not become a dominant trend.

At present, the variety of topics published in the journal is very impressive. Based on the classification by topics provided by the Index to Survey Methodology volumes 1 to 24, out of 420 articles, 31% dealt with estimation, 20% with sampling design and survey development, and 12% with non-sampling errors. The other articles dealt with analytical methods, data collection, quality assurance, edit and imputation, confidentiality and a few with general topics. It should, perhaps, be encouraged that papers on questionnaire design and non-response, the weakest links in survey design, be more frequently published in the journal.

For the past several years, the journal has been extremely ably guided by Gordon Brackstone, Chairman of the Management Board, M.P. Singh, an excellent Editor since the beginning, and Frank Mayda as Production Manager. The Editor is assisted by Associate Editors, who come from universities, government agencies, and private sectors around the world, ensuring a desirable mix of theoretical and practical interests. The large number of subscribers from many countries (70) gives a clear testimony to the journal's broad appeal and importance. Statistics Canada should be proud of its journal. The journal is not only effective publicity for the organization, but is also a leading methodological publication in the world.

REFERENCES

SINGH, M.P. (1988). Survey Methodology. In *Encyclopedia of Statistical Sciences*, (S. Kotz and N.L. Johnson, Eds.), 9. New York: John Wiley and Sons.

STATISTICS CANADA (1975). Editorial Policy. *Survey Methodology*, 1.

STATISTICS CANADA (1984). Editorial Policy. *Survey Methodology*, 10.

STATISTICS CANADA (1988). Survey Methodology Management Board minutes, July 1988.

WILK, M.B. (1982). Letter to R. Platek.

ZARKOVICH, S.S. (1985). Letter to R. Platek.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 113-128
Statistics Canada

Statistical Services – Preparing for the Future

IVAN P. FELLEGI[1]

ABSTRACT

In this last year of the 20th century I propose to look forward: to the challenges facing statistical agencies and how to prepare for them. I will first of all review the context within which statistical offices must evolve: first the main forces at work that are modifying our economy and society; second, specific policy issues that need to be addressed; and third, changes in the nature of government and in expectations from it. I will then try to outline an internal strategy for statistical offices derived from the foregoing analysis. On the one hand, this will require the development of new types of data systems that are needed. I will illustrate these with reference to recent initiatives by Statistics Canada. While crucial, these new data systems will probably only account for a relatively small proportion of our expenditure. Hence a second important component of our strategy designed to cope with external social and economic changes must be to ensure a high level of adaptability of our core activities. Such adaptability requires specific initiatives. Finally, I will describe the elements of an "external strategy": how to manage our interactions with the world around us. Three elements will be touched upon: achieving and maintaining a high level of relevance; the issue of political objectivity and its perception; and international collaboration. In each case I will try to outline specific measures that I consider essential.

KEY WORDS: Statistical organization; Strategy; Planning; Conceptual frameworks; Relevance; Non-political objectivity.

1. INTRODUCTION: THE POLICY CHALLENGES

An early version of this paper was prepared for the 1997 UK Statistics Users Conference (London, November 1997). It also formed the basis of the Gold Medal address delivered to the 1998 Annual Meeting of the Statistical Society of Canada.

Peering into the future and discerning relevant trends is never easy. Even more difficult is the derivation of a strategy for statistical offices.

Indeed, there is a saying that "forecasting is very difficult – particularly of the future." I am therefore very fortunate that, by coincidence, I can rely on others for part of my job. Recently the head of the public service of Canada (the Clerk of the Privy Council) asked a group of senior policy analysts from major departments (I will refer to them as the Policy Group) to produce a paper "on the pressure points that are likely to arise in Canadian society by the year 2005 as a result of economic, demographic and social trends". Statistics Canada played a major part in the events leading up to the commissioning of this report and in its preparation. The Chief Statistician was asked to chair an interdepartmental committee of senior officials to report on the current capacity for policy analysis of the government. One of its key recommendations was that such capacity is maintained in response to demand from the Cabinet and the clerk of the Privy Council for serious policy analysis. A specific follow-up action was the commissioning of the report from the Policy Group. The report drew extensively on an in-depth analysis prepared by Statistics Canada on important long-term trends. They reported in October, 1996 and I will start with an outline of their conclusions. While the report was written from a Canadian perspective, I think their findings have relevance for most developed countries.

Next, I will touch upon some trends in the political environment and governance that are of relevance to statistics. The bulk of the paper is devoted to an outline of what I think might be a robust strategy for statistical offices.

2. MAJOR POLICY CHALLENGES

2.1 Main Forces at Work

This part of the paper is based on the report "Growth, Human Development, Social Cohesion" prepared by an interdepartmental committee for the Clerk of the Privy Council Office, Canada, in October, 1996.

The Committee identified five forces at work with pervasive impacts on a broad range of policy domains.

2.1.1 Globalization

Globalization is the integration of production and distribution across national boundaries. In response to dramatic declines in shipping costs, customs barriers and the phenomenal evolution of computer communications, multinational companies and complex partnerships can operate as integrated entities, even though their operations span the globe. This enables them to exploit the relative advantages of each location with little loss to the traditional benefits of tight centralization. What is happening is not just a phenomenon of increasing exports, but altogether new production and distribution arrangements that render national boundaries increasingly irrelevant. Globalization affects not only the economy but all aspects of society and culture. Indeed,

[1] Ivan P. Fellegi, Chief Statistician of Canada, Statistics Canada, R.H. Coats Building, Ottawa, Ontario, Canada K1A 0T6.

a major challenge will be to obtain the benefits of economic integration, while maintaining independence in other domains, such as social programs, culture, and the environment.

2.1.2 The Information Revolution

The speed of change of the so-called information revolution was well captured by The Economist (1996), "If cars had developed at the same pace as microprocessors over the past two decades, a typical car would now cost less than $5 and do 250,000 miles to the gallon". This is the development at the heart of the globalization of economic production which affects the delivery of services perhaps even more fundamentally than that of the production of goods. All sectors have made massive capital investments in information technology, but so far – and if we trust our productivity measurements – society's return on these investments has been limited. Nevertheless, there is continued hope for major future productivity improvements, as a result of the information revolution, with a consequent and sustained improvement in economic performance.

However, technology will affect different people differently. So there is also a major risk that a new and pervasive social fault line might evolve, one which divides those with a mastery and access to technology from those who are ill equipped to do so. Technology also affects culture and social cohesion and it promotes the formation of interest groups which transcend national boundaries. Identification with traditional unifying entities such as nation or city have come increasingly under pressure.

2.1.3 Environmental Pressures

We are all aware of a series of environmental pressures, such as the threat to the ozone layer, global warming, or the collapse of fishing in certain traditional areas. But our awareness is not matched by a thorough understanding of either the full risks, or of the linkages between changes in economic or social activity and the environment. This makes it particularly difficult to weigh potential trade-offs among economic, social, and environmental objectives.

2.1.4 Various Demographic Pressures

Most developed countries have a birth rate that is below the replacement level. Combined with continued progress in longevity, this results in societies in which the weight of the aged is increasingly felt. The trend will accelerate, at least temporarily, when the post-war generation of baby boomers reaches the retirement age.

The continued evolution of the family as the basic social unit poses a series of ill understood challenges. Not only has the role of women changed as the two-earner family became the norm, but different forms of family are gaining in prevalence (single parents, common law, same sex couples). The impacts of these changes are not fully understood but are certainly felt in such important policy domains as poverty, labour market activity, child development, pension plans, and caring for the chronically ill elderly.

Immigration has become, not only for Canada but for most of the OECD area, the dominant source of population growth. While it is generally regarded as a positive factor, the integration of people from a variety of cultures has implications for social norms, policy and collective identity. Illegal immigration, somewhat less of an issue in Canada, is a major source of current problems and a potential source of significant future social divisions in many developed countries.

These developments, and others, have cumulative impacts that extend well beyond the present time. The evolution of a child into a happy and productive citizen, the development of cancer or heart disease, the availability of a pension at the time of retirement are examples of processes that extend over long years and involve complex interactions. The supreme policy challenge is to gain a better understanding of causal factors, particularly those amenable to modification through social programs. Much better information can be of immeasurable help in gaining the necessary understanding.

2.1.5 Fiscal Context

All the pressures described above, and the policy space available to deal with them, interact with the fiscal context, *i.e.*, the balance to be struck between reducing the accumulated debt of government versus the resources to be made available to deal with current or emerging issues.

2.2 Policy Challenges

The Policy Group identified a number of specific policy challenges, many of them closely interwoven with the pressures outlined above. They grouped the challenges under three broad headings, of which I will touch on two: growth, and human development. I can only give here a brief indication of what they saw as areas of particular concern in Canada.

Growth

- *Economic growth.* The growth of real GDP per capita, as well as of real average family income has slowed considerably in the last twenty years from what it was in the previous decades.
- *Productivity.* In spite of the wide scale introduction of computer technology during the last twenty-five years, productivity growth rates have declined in several of the major industrialized countries. The reasons for this are ill-understood, in spite of the crucial importance of productivity for economic growth.
- *Adjusting to the so-called knowledge-based economy.* We are aware of a shift in the economic centre of gravity towards services and knowledge and technology-intensive industries. In terms of both employability and income, there are considerable and still

increasing returns to individuals from education, and perhaps even more from literacy and life-long learning. Similar findings apply to firms that innovate and adjust. But, once again, we have a very inadequate understanding of the relationships among such forces as individual skill acquisition, the business practices of firms, the growing shift to non-standard forms of employment, the use of technology, innovation, and productivity.

- *Environmental issues.* These interact with the economy, as well as with most other domains. There are basic questions, very poorly understood, that are loosely described under the heading of sustainable development.

Human Development

- *Imbalances in time use.* The Policy Group drew attention to a number of what they called imbalances in time use.

  Less time is used working, compared to being in retirement. The ratio of one to the other has declined by a factor of two in thirty years as the age of entry into labour force continued to increase even while the age of retirement continued to decline. This development, which appears to be part of a longer term trend rather than a simple consequence of the business cycle, places substantial pressures on both public and private pension systems.

  The distribution of available work is getting more polarized: in the last twenty years the number of persons working both short and long hours has steadily increased.

  Large scale unemployment coexists with a serious "time crunch" experienced mostly by young working couples with children – with potentially important but ill-understood implications for the quality of care received by their children and their elderly parents.

- *Labour market issues.* There are important visible signs of major changes in labour markets. Unemployment in most industrialized countries has increased from one economic cycle of the post-war period to the next; more jobs are non-standard (part time, temporary and self-employed); individuals change their status more frequently between employment, unemployment and spells of education; there is a polarization of both earned incomes and hours worked; education and perhaps even more so skills are becoming increasingly important. While we regularly monitor these changes, the underlying forces involved are poorly understood.

- *Transitions.* Besides those transitions occurring within the labour market, other key changes in life are from home to school, school to work, childlessness to parenthood, and from work to retirement. We have very little quantified knowledge about the factors that result in successful transitions, or about the interaction of the various forces at play.

- *Increased polarization of incomes and the role of the tax/transfer system.* A significant finding is that, in Canada as in the United States, the polarization of gross family incomes has increased over the last two decades: those with high incomes increased their share in the total while those at the bottom of the income distribution have reduced theirs. Unlike the United States, in Canada this trend was fully offset by the progressive character of the tax/income transfer system – at least until now. There are important policy questions raised by these trends about the role of social safety nets, particularly in a period when a fight against government debt has a high priority.

- *Children and youth.* One of the most difficult issues relates to youth employment. The unemployment rate for young people is stubbornly high. Furthermore, this has persisted in the face of more youth staying in school longer, thus alleviating somewhat the pressure on the youth labour market. Even when they are employed, the real earnings of young workers have declined over the last twenty years, in contrast to that of their elder colleagues. Our understanding of causes and possible remedies is very incomplete. There are issues related to children that are even less well understood, *e.g.*, the impacts on children of the socio-economic status of parents, their parenting styles, the education system, the teacher, the neighbourhood, the environment, or heredity.

- *Health.* Prompted by government expenditure controls, considerable public debate is focussed on cutbacks in hospital and physician services. Yet, there is mounting evidence that the formal health care system is only one of the factors affecting population health, and perhaps it is not even the most important one. Some of the others are life-style, socio-economic status, the environment. Even within the formal health care system, there are well documented and widely divergent practices (for example in the propensity to use coronary bypass operations or Caesarian sections), with important cost implications, but whose longer term impacts on health are poorly understood.

- *Aboriginal people.* A whole series of policy issues relate to the aboriginal population – a major issue in Canada. On almost any socio-economic scale they continue to score considerably worse than the rest of the population.

- *Law and order.* Opinion surveys show high and still growing public anxiety about crime. Perhaps in response to public anxiety, the rate of incarceration is growing much more rapidly than any measure of crime – even though there is no evidence that more incarceration results in less crime.

The list above is not exhaustive, nor does it need to be. I hope it suffices to indicate the broad concerns which dominate our public policy agenda. I believe it also demonstrates that, while our current statistical system largely succeeds in describing the phenomena of concern, more often than not it does not provide nearly sufficient help to understand them.

3. TRENDS IN GOVERNANCE

Without any doubt the single most important role of statistical agencies is to assist the public policy process. Consequently the role that governments play, and the public expectations regarding such a role, have a crucial relevance for us. It is therefore worth reviewing briefly recent trends and likely developments.

During much of the post-war period, the economic concerns of the governments virtually everywhere in the OECD area were macro-economic, and their attitude was interventionist. The statistical response to interventionism in fiscal, monetary, industrial and trade related affairs was to develop a comprehensive system of economic accounts supported by an equally comprehensive system of business and household surveys.

In the social domains, the post-war years saw the establishment of great universal programs in the fields of health care, access to post-secondary education, unemployment insurance, pensions, and so on. These programs imposed relatively less stringent demands on the statistical system. We were asked (perhaps) to measure their cost, other major inputs (*e.g.*, health and education personnel), key operating ratios (students per teacher, population served by each physician), "raw" measures of output (students graduating, number of people discharged from hospitals), and so on.

The situation has changed – gradually in the eighties, and much faster in the nineties. A number of factors were responsible for these changes. Slower economic growth, apparently ineffectual macroeconomic policies, and the cost of universal social programs which almost invariably exceeded the initial estimates, resulted in mounting deficits. At the same time, globalization of financial markets increased the pressure on governments to "deal with the deficit". The issue, therefore, rose to the top of the national agenda, implying substantial cuts to established programs that necessarily lead to questions of "what works", and "for whom is the program essential". The statistical system was not well prepared to answer these difficult questions.

The retrenchment of government programs is always a politically difficult task. It was particularly so in the midst of slow growth and persistently high unemployment. The combination exacerbated the concerns of the public – already cynical about the role of government and indeed about government itself.

All these developments lead to a search for evidence about the real impacts of government programs on society and the economy. The interest was spearheaded by governments themselves who wanted and needed information for their own analysis prior to making controversial decisions about the elimination or major modification of established public programs. But they also needed new and detailed information as objective support in public debates. Indeed, a new movement emerged requesting from government that it identify and officially adopt performance indicators with a focus on outcomes, as opposed to processes. The combination of increasing intellectual rigour and limited financial means to initiate major new programs has actually lead to a view according to which it is a core function of governments to ask the right questions, and to ensure the availability of information needed to answer these questions.

These developments clearly have major implications for statistical offices. In the economic domain a high priority continues to be placed on the monitoring of macroeconomic developments. While maintaining their macro-economic interest, recent Canadian governments have been paying much greater attention to microeconomic considerations, *i.e.* understanding the factors that account for successful business outcomes and attempting to underpin macroeconomic intervention with coherent policies designed to help business at the micro level.

Daunting as this task is, the new requirements are even more demanding in the social domain – and for several reasons. First, outcomes are typically the results of long term effects, for example in the health and education domains. These cannot be traced back unambiguously to unique causes. At best one can hope to identify factors that tend to move outcomes in different directions. Furthermore, policy interest shifted from considerations of broad social impacts to the examination of impacts on particular groups – an intrinsically difficult task. And to make the task even more difficult, the weighing of policy alternatives unavoidably involves an examination of their expected future impacts (a form of simulation). In effect, we are asked to provide the rich statistical data base needed to identify the policy levers which, at an acceptable cost, are likely to result in some specified desired outcomes (Fellegi and Wolfson 1997).

Even as official statistical agencies are called upon to become much more policy relevant, they must do so in a manner that preserves both their political independence and reputation for professional competence. Indeed, this has become more important than ever before, precisely because of the heightened impact of official statistics, combined with an environment in which the government is trying to convince a sceptical and insecure public of the wisdom of its actions, making use of evidence that must be accepted as "objective" (Fellegi 1991a).

4. IMPLICATIONS FOR STATISTICAL OFFICES: AN OVERVIEW

One can argue whether the Policy Group correctly identified the forces that are most likely to have a large impact on our evolving society, and whether they correctly deduced the particular areas of policy which will require the greatest attention. I submit that this would not be particularly productive. After all, societies do not have a particularly good record of forecasting the challenges that will face them in a few years' time. Nevertheless, there are some important implications for us that we ignore at our peril.

Even though I would not propose the uncritical acceptance of their specific list of expected policy challenges, I think that it is possible to identify broad domains in which major challenges can be anticipated (the issue of unemployment, the evolving character of employment, productivity, the causes of health and illness, the role of education and training, *etc.*). After all, we are developing major statistical systems to illuminate a broad policy area, and not to support some specific policy initiative. The domains, together with the nature of the information that is likely to be needed, should give us considerable guidance regarding the statistical developments that are needed.

It is clear from the work of the Policy Group that many of the areas that they identified are characterized by fundamental gaps in understanding. Garnet Picot (Picot 1997), for example, analysed a variety of government measures designed to deal with the issue of high unemployment and showed that their likely effectiveness depends on which of several possible causes we think are primarily responsible for the problem. There are a number of plausible contributing causes: low economic growth, disincentives to accept low paying jobs (due to the character of social transfer systems and unemployment insurance), high payroll taxes which render hiring of new employees more expensive, a mismatch between skills needed and available, high minimum wages, *etc.* Disentangling their relative importance is a prerequisite of sound policy remedies. In turn, he describes the statistical data systems, some of them quite innovative, that are needed to accomplish this goal.

I would also take as a good working assumption that the political environment will continue to evolve broadly along the lines discussed in the section on trends in governance – with clear implications about the general character of the statistical service that the country needs.

In summary, therefore, we can foresee major challenges. In some instances we must improve existing data systems, and in others start monitoring phenomena whose priority has increased (*e.g.*, the environment). I expect that the greatest challenge will be to devise new data systems designed to go beyond monitoring and with an explicit orientation to shed light on the underlying dynamics. While we might well be able to identify the domains in which such information will be needed, there will be considerable uncertainty about the precise policy issues which we will be called upon to illuminate – with great insight but also with great objectivity. In the face of these challenges our strategy must be bold, but also robust – hence evolutionary. The three ideals that we must pursue are: relevance, adaptability, and objectivity.

These three themes will be the themes of the rest of the paper which will be divided into the following sections: prerequisites and approaches leading to essential new data systems required to illuminate policy domains and to inform public policy discussions; strategies with respect to our core programs and the organizational flexibility needed for their further evolution; and finally a section devoted to what might be called an "external strategy" to ensure and safeguard both program relevance and non-political objectivity.

5. IMPLICATIONS FOR STATISTICAL OFFICES: NEW DATA SYSTEM

5.1 Elements of a Strategy

It is at the heart of our challenges to try to develop the new types of statistical systems that are needed to inform public policy discussions in key fields. Policy development is a complex process. It involves politicians, the public, special interest groups, as well as researchers and policy analysts within and outside government. Experiences and views of all these groups, as well as political ideology, combine to arrive at policy proposals. It is essential, however, that there should be relevant empirical and theoretical evidence to both nourish and to temper the views of all participants. What is needed is a concerted effort to try to understand the forces at work, to be able to anticipate with confidence the likely performance of alternative policy levers. This does not usurp the role of the political process, but rather helps to inform it (Fellegi and Wolfson 1997).

We must be clear, however, that helping to understand economic and social phenomena, as opposed to monitoring their impacts, is a task that is at least one order of magnitude more difficult. In this section I will outline the elements of a broad strategy that we have attempted to follow in Statistics Canada. I will then provide some illustrative examples of new data systems implemented at Statistics Canada.

(i) Development Must be Rooted in Relevant Outcome Measures

A key end objective is to help the public policy process to discern the relevant "policy levers" that are likely to be most effective in moving us toward the achievement of desirable social and economic objectives. It is logical, therefore, that the development of new data systems should start with an attempt to understand the key outcomes that policies in a given domain would like to promote. This does not imply either policy advocacy or a politicization of the statistical system: it is in the common interest that the goals of government be carried out on the basis of the best available information, and that the same information be made equally available to others who might wish to promote different policies.

A source of major difficulty is that in many areas, particularly in the social domain, there are no broadly accepted outcome indicators. For example, while we all agree on the objective of wanting to-improve education, there is no general consensus on what this should involve and how progress should be assessed. Do we want the outcome of a good education system to be people who are successful in the labour market, who are well rounded individuals, who have acquired the skills to continue to educate themselves in an adaptive mode, who embrace the values of good citizenship, or some combination of the above? Difficult though these questions are, they are clearly essential to the development of a truly relevant and useful system of education statistics.

(ii) Connecting Outcomes and Policies is a Major Objective

High quality and operationally measurable outcome indicators represent the first step in developing useful data systems. They are fundamental to monitoring progress toward objectives, telling us whether or not improvements can be detected. But good outcome indicators alone are insufficient since, at least in democracies, society's ultimate objectives seldom lend themselves to direct intervention: *e.g.* sustained employment cannot be generated directly, nor can the health of the population be improved by any direct measures. Truly useful statistical systems must therefore allow observers and analysts to discern the relationship between outcomes and public policy interventions: the so-called policy levers.

We cannot assume that the traditional policy levers are necessarily the most important ones. For example, there is growing recognition that class size or pupil-teacher ratios do not have a particularly large influence on student outcomes. Similarly, health may have as much to do with family income as with the medical interventions of the health care system. The implication is that policy relevant data systems should be based on broad views of the relevant causal factors, and the ways they relate to each other – in other words, they should be based on a conceptual framework.

(iii) Conceptual Frameworks are Prerequisites

There is no widely shared agreement about what constitutes an adequate conceptual framework for a statistical system. A conceptual framework is not, in and of itself, a theory. It is, rather, a carefully constructed reflection of our present understanding of the forces at work which have a potentially significant impact on our selected outcome indicators. I believe that one of its essential features has to be the attempt to reflect, at least schematically, the forces at work within a given domain, including their interactions and the direction of their effects.

A useful conceptual framework is neither abstract, nor static. It gains both its empirical and adaptive usefulness from being dynamically coupled to a measurement system. On the one hand, conceptual frameworks should guide the evolution of those data systems which quantify the interactions displayed by the framework. On the other hand, data systems should have a profound influence on the evolving conceptual framework: they should lead to the elimination of insignificant or irrelevant relationships (for example, of ideological dogma!), and to the further elaboration of those that are most important. Furthermore, data analysis might bring to light altogether new relationships and insights and in due course may lead to a revision of the framework.

(iv) Partnerships are a Necessary Condition for Progress

The type of statistical evolution outlined here represents a very ambitious undertaking. The effort is justified by the overwhelming need to improve our understanding of the forces at work underlying the most pressing and vexing social and economic problems. This is clearly a necessary condition for more effective policies and programs. The cost, while not negligible, is dwarfed by the cost of government programs designed to deal with ill-understood problems.

The major effort that is required cannot succeed without an intensive tripartite collaboration involving the sectoral policy department concerned, the statistical office, and the social science community. The policy department's support must take at least three forms: moral support, direct financial support (or alternatively, support for the financial requirements of the statistical office), and effective collaboration in exploiting existing models and data systems in the course of analysing existing government programs and policy options.

The participation of the social science community is fundamental to ensuring that prevailing theories are brought to bear, helping to design the instruments for shedding light on the relevant phenomena, testing and modifying theories. Their analytic work should play a critical role in highlighting missing statistical information needed to test existing theories. At the same time, social science benefits from availability of quantitative information which can be used to sort out idle speculation from empirically validated hypotheses.

Of course, the role of statistical offices in such tripartite collaboration is highly significant. They must take the lead in convincing decision makers of the critical importance for them – the decision makers – of launching the long term developmental process that the provision of the right information at the right time requires. They must also take the lead, with help from policy analysts and social scientists, in identifying and, if funding can be secured, implementing the needed data systems. In our experience, these systems tend increasingly to be longitudinal surveys or administrative systems in domains such as health, education, and labour market and income. Longitudinal data, to a much greater extent than cross-sectional, can associate outcomes with a range of possible causal variables – a prerequisite for the identification of causal links.

It is fundamental that the required extensive collaboration occur and that it be productive. The leadership drive to start the process may come from anywhere – including, I should emphasize, official statisticians as well. We do not need to wait for others to come to us for our help. On the contrary, we should have a well developed analytic program on the basis of which we are able to articulate not only what are the data gaps, but also what are the areas of public policy where so-called "evidence based decision making" is not possible because of missing conceptual frameworks and supporting data systems.

Statistics Canada, in collaboration with our government and academic partners, did take a number of significant initiatives to develop data systems designed to lead to significantly improved understanding in several important public policy domains. In a subsequent part of the paper I shall briefly describe a few examples.

(v) Practical and Useful Approaches are Sectoral

In spite of the increasing recognition of the interrelatedness of social and economic phenomena, there are both substantive and practical reasons for preferring the development of new data systems along either sectoral or functional lines. There are separate ministries and other organizations dealing with policies in the fields of human resources and labour markets, trade, industry, welfare, health, education, justice, *etc.* Data systems can only become policy relevant if there is a constituency for the information, one whose function is to consider the implications of the findings. But beyond this "mundane" consideration, it is almost always the case that outcome measures are formulated as sectoral (The word "sector" is used here to connote a domain which is the subject of particular government policies and programs: health, education, labour markets, macroeconomic policies, and so on. Typically, a government department is assigned either sole responsibility or at least a lead role in respect to policies and programs in that sector.) objectives: to reduce unemployment, improve population health, improve the effectiveness of the education system, and so on. If the data systems are to shed light on our performance, they must specifically be developed to improve our understanding of the given sector (Fellegi and Wolfson 1997).

We know that there are important interactions among the sectors: for example, both income and education are known to affect health, while also health affects education. But these effects can be accommodated within the sectoral models as exogenous variables.

5.2 Examples of New Data Systems in Statistics Canada

Statistics Canada is engaged in the development of new data systems in a number of domains, broadly in line with the areas of policy challenge identified by the Policy Group. A common characteristic of these initiatives is the goal to illuminate a given policy domain, *i.e.*, to identify the main forces at work and to measure their relative importance. Where a broadly shared understanding exists about what these forces are, the main task consists of quantifying their respective strengths in affecting the outcomes of interest. This is the case, for example in the area of income and labour market dynamics. In other instances a major initial effort is needed to outline a conceptual framework which is subsequently fleshed out with data and modified through use. This is the case with health, and science and technology. In the case of education and child development we could only identify what we, in collaboration with our partners, thought is a reasonably comprehensive list of all possible forces (but without the complex interactions among them). In this case we decided to start with a very comprehensive survey, hoping to elaborate a more complete conceptual framework through the analysis of the resulting data.

Most of these initiatives take the form of longitudinal surveys. This is not surprising since cross-sectional surveys can monitor phenomena, but only longitudinal surveys are capable of linking outcomes to their correlates – an essential prerequisite for the analysis of the relative importance of alternative "policy levers" and other (exogenous) variables.

(i) Labour, Income and Family Dynamics

In the socioeconomic domain the single dominant problem facing most of the G7 countries is the persistence of high unemployment, the attendant poverty, and the possibility of a semi-permanent underclass (those caught in the "poverty trap"). More generally, the issue is the relationship between income, labour market participation, and personal as well as family circumstances. There are a number of important questions: under what circumstances do poor families manage to escape poverty? What personal characteristics and what government programs appear to help single parents to cope successfully? Under what conditions do unemployed youth, particularly those without post-secondary education, manage to break out of the vicious cycle of "jobs require experience but experience can only be acquired on the job"? What factors account for successful and unsuccessful transitions from school to wórk, from job to unemployment, from work to retirement?

In Canada we have started an on-going program to shed light on this complex of issues. It is based on a sample of families, each of whose members will be tracked for at least six years as they move through various labour market and income experiences, as some members move out of the original family and perhaps form new families, and so on. Key characteristics of the program are its longitudinal dimension and the explicit objective of trying to link causes and effects; the close collaboration in its development between Statistics Canada, the main policy department, and the academic community in the conception of the survey and its analysis; and the explicit objective of maintaining a capacity for the further evolution of the survey. It was felt that, in this domain the existing literature had already identified the main forces at work, and what was needed were suitably organized data on how these forces play out themselves in Canada.

The survey includes core questions on incomes, labour market experiences, and family characteristics, but there is also room for supplementary questions to explore new hypotheses that the analysis of the data might suggest.

(ii) The Interaction of Business Performance and Employee Outcomes

There is increasing suspicion that productivity outcomes can only be studied at the micro-economic level and that a number of business practices may be ultimately related to output per unit of labour. For example, at what rate and with what effects is information technology used in the workplace? Does the effective use of technology imply a higher investment in skills upgrading? Do employees with lower level skills risk becoming "disposable"? Is the use of flexible labour market practices (increasing use of contracting out and of temporary or contingent workers) a significant contributor to business success?

We have carried out a pilot to establish the feasibility of an on-going longitudinal survey of business establishments, including a subsample of their employees who would be tracked for at least one extra year after their employment with the selected business comes to an end.

The survey we expect to take would provide information on the extent of use of new technologies, on training available to employees, business strategies pursued (*e.g.* the extent and role of R&D; the emphasis on new products; expansion into new geographic markets; collaboration with other firms in R&D, in production, or in marketing; *etc.*), labour market strategies (*e.g.*, downsizing, re-engineering, greater reliance on part-time or temporary workers, increased use of overtime in place of new hiring), degree of market competition, change in the occupational composition of employment. The information collected in this fashion will be related to "objective" business performance indicators such as value of production, sales, exports, profits, *etc.* It will also relate the firm's behaviour to impacts on employees: their training and the acquisition of new skills; the relationship between the use of technologies and wages; training and other worker outcomes; and the relationship of all these factors to employment stability.

(iii) Survey of Children

We have an incomplete understanding of education and child development, even more so than of the dynamics of incomes and employment. What are the key influences that lead to the development of productive and happy members of society? My third example involves a very ambitious longitudinal survey of children, initially of 0 to 11 years of age, which attempts to shed light on this rather basic question. Because of clear indications from existing research that causal factors in this domain operate over very long time periods, the objective is to follow the same sample of children well into young adulthood – up to 20 years. However, the survey is arranged to provide important analytical and policy-relevant results on a continuing basis. We are collecting a wide range of possibly relevant explanatory variables related to them: mother's health during pregnancy, socioeconomic conditions of the family, parenting styles (the nature of parent-child interactions and stimulation), early signs of emotional or learning problems, socialization (relations with peers and potential friends), scholastic tests, teacher's assessment of the child, and the principal's assessment of the school.

In this case we did not have a fully developed conceptual framework to guide our survey development. However, working in the closest possible collaboration with relevant researchers and academic staff, we could identify a long list of variables which could have a material impact on child development. This resulted in the exceptionally wide range of variables that are included in the survey. Rather than starting with a fully developed conceptual framework, we plan to approach its refinement through analysis of the survey data. Indeed, we have arranged a wide range of contracts and other forms of collaboration to ensure a full exploitation of the data base.

As in the case of the previous examples, the survey can be adapted from one round to the next to reflect our gradually improving understanding, or simply to collect some additional information in a cost effective manner.

(iv) Population Health

Most G7 countries spend on health 7 to 10 per cent of their GDP. Health is also a policy area that is consistently near the top of the list of greatest concerns to Canadians. Yet, here again, the substantive policy challenges far outstrip our ability to provide information that would support "evidence-based decision making" regarding the determinants of population health and the long term impact of health interventions. Our third longitudinal survey is designed to gain some insights in this domain.

The survey follows a sample of individuals for a period of years yet to be determined. It contains a core set of questions, including those on health status, disability, health care utilization, health problems, family situation, and labour market participation or other major activity. In addition each survey also contains a series of questions that delve into a specific topic for that cycle only (the initial cycle focussed on mental health). There is also an arrangement for linking respondent records with provincial records of health care administration in order to incorporate into the data base the encounters of sample persons with the formal health care system.

Prior to the development of the survey we did invest considerable effort to develop an explicit conceptual framework to guide us. As in the other examples, the design of the survey instrument and the supplementary inquiries incorporated into each round have been developed in close and continuing collaboration with the main federal and provincial stakeholders of the health field, as well as with a broadly representative group of advisors.

(v) Science and Technology

The need to understand the impact on society of science and technology has risen high in our policy agenda. Much has been written about the importance of adequately investing in science and technology – but much of it is unsubstantiated. Is it true that there is a close link between a country's investment in science and technology and its rate of economic growth? What are the impacts on employment? On the physical environment? On social cohesion? What is the relative contribution of different sectors of society (government, university, business, *etc.*) to the generation of knowledge and what are the results? What can we say about the storage of knowledge (both informally, in people's heads, and in formally accessible devices such as books, diskettes, *etc.*)? How is it used, with what effects? How is knowledge generation financed?

In order to attempt an answer to these questions we are proceeding along several tracks. With the help of an external advisory committee, we began to develop a conceptual framework. As this framework is defined, we are reviewing existing information to assess which parts of the framework it supports and where are the important data gaps. At the same time, we are trying to understand the main policy questions which ought to be answered. This will allow us to outline a multi-year program of information development in order to improve our understanding of key questions affecting our economy and our social organization. And even while the conceptual work is proceeding, we are beginning to collect relevant information that will clearly be needed: on innovation, on the adoption of advanced technologies, and on knowledge flows to business from universities.

(vi) Productivity

My final example deals with the issue of productivity. Collectively, we are investing heavily in the new technology embodied in computers and related telecommunications and software. The September 28, 1996 issue of The Economist estimates this combined investment at 12% of total capital stock, the same as the level of investment in railways at the peak of the railway age in the 19th century. Normally, when heavy investments are made in a new technology, one expects significant productivity returns. Yet in most G7 countries there is a decline in the measured rate of productivity growth during the last twenty years compared to the previous twenty. The Economist calls this finding "a statistical black hole".

Many believe that, to the extent there is a measurement problem, it has to do with the measurement of the output of several high growth service industries: banking, telecommunications, consulting, and so on. The problem in these sectors is that it is hard enough to define the unit of output, let alone the quality improvements to which these outputs are subject at an increasing rate.

This area also provides an example of effective international collaboration. Statistical offices of several countries decided to work together, drawing on the relative strengths of each. In fact, their representatives agreed to develop a conceptual framework and corresponding "model surveys" for particular service industries, often in collaboration with leading businesses in their respective countries. These model surveys have been piloted in other volunteer countries and experiences compared. Not only are improved measurement techniques developed in this manner, but as a byproduct international comparability of data is also achieved.

6. IMPLICATIONS FOR STATISTICAL OFFICES: ADAPTABILITY AS KEY STRATEGY

6.1 The Core Program

Even while new policy domains gain prominence, most of the problems of the postwar years continue to be relevant. Undoubtedly, we shall continue to collect data on economic growth, inflation, employment and unemployment, the evolution of incomes, education levels, health, *etc.* – phenomena whose monitoring remains intrinsically important even if our level of understanding of how to improve our performance in each of these domains is relatively limited. We do not stop taking the temperature of the patient just because we do not fully understand the reason for the fever, or how to cure it. For example, we may not have succeeded in understanding how to cure unemployment, society learned to make a variety of accommodations to it. Some of these involve adjustments of individual behaviour. Others are more programmatic and try to alleviate its worst effects (*e.g.*, unemployment insurance), or try to bring about improvements (*e.g.*, labour market training programs). We may adjust the patient's diet, try to bring down the temperature to avoid secondary complications, and so on.

While our core function will certainly continue, we can expect important challenges – both old and new.

First, the challenge of finding the right mix between preserving continuity and adapting our concepts to changing reality will continue to be with us. Here, as in so many other places, international collaboration is needed and productive. Indeed, in the face of conceptual difficulties we can speed up progress by pooling our respective strengths. Furthermore, where the conceptual underpinnings are weak, international practice and convention confers a degree of legitimacy. International standards are all the more important since an increasing proportion of our clients operate in a global context and want information on the basis of which valid comparisons can be made of the performance of different countries in different domains.

Second, this being our core program, it is where the bulk of our expenditures are spent, and where we must look for efficiencies. And we will certainly need to find savings since we are inevitably called upon to contribute to the funding of new statistical initiatives. In my experience, necessary conditions to obtain new government funding are to be seen as efficient in carrying out our core program, and to be able to show that we have identified and eliminated programs of lower priority. But to do so we must have an effective planning system (Fellegi 1992).

The third challenge with respect to our core activities is to improve our dissemination program, particularly as it affects the needs of the general public, most of whom receive their statistical information via the media. We have to do much better in informing them about findings, as opposed to releasing data. Emphasizing findings as opposed to releasing data has an enormous impact on how the public perceives our relevance (Fellegi 1991b).

6.2 Ad hoc Survey Capacity

The need to conduct special surveys arises when a client requires some information which cannot be met from the regularly funded program of the statistical agency. The capacity to respond to ad hoc requests, provided they are accompanied by the required funds, represents an important form of flexibility/adaptability of the statistical system. The requirement may take a number of forms, but I will restrict my attention to surveys or pilot surveys. As I indicated in (Fellegi 1996), there are compelling reasons to maintain a strong capacity for client funded surveys.

Special surveys result in new information being placed in the public domain, often in new areas. Charging for the development of ad hoc information provides a type of market test: if a contracting department is willing to spend significant funds from its own budget, then the resulting information is likely to be relevant to serious policy concerns. Since all official statistics, irrespective of the source of funding, should always be available to the general public, the external funding of policy relevant information is clearly in the public interest (While a strong capacity to respond to the ad hoc requirements of client departments is important and unambiguously beneficial, there are strong reasons of both statistical policy and efficiency to prefer that the regular statistical needs of clients be met from the regular budget of the statistical office.).

Additional benefits of having a capacity for client funded surveys include the following:

- client-sponsored surveys are safety valves for demand which cannot be satisfied within the budget constraints of the statistical office. The flexibility to respond, therefore, has a major.impact on client satisfaction with the statistical system;
- special client-sponsored surveys typically relate to new areas and hence involve innovation. As such they contribute importantly to the maintenance of an environment that is open to new ideas. Indeed, some of our cost recovered surveys are pilot surveys designed to test new approaches prior to seeking regular funding for them;
- to the extent that charges include full costs, including overhead, they contribute to the maintenance of an operational capacity;
- Statistics Canada has established two divisions, both capable of rapid expansion whose budget derives entirely from client-sponsored surveys. The staff work very closely with major client departments and have learned to "market" not only their operational capacity, but also their ideas. These divisions necessarily evolve a culture of client orientation whose benefits are far reaching.

How to create and maintain a special surveys capacity? After all, to a limited extent every statistical office is able to carry out one-time work by mobilizing the needed resources. However, I think that we need to go well beyond that. We need to create an organizational culture that welcomes, indeed seeks out such work – otherwise it risks becoming an extra chore accepted grudgingly. The principal means used by Statistics Canada to achieve this broader objective were:

- the creation of the two divisions mentioned above, which operate respectively, in the households/social domain and in business statistics. Both entities have to "earn their keep" through contract work;
- a marketing orientation for the entire organization, including explicit revenue targets;
- a personnel and financial management environment that encourages the deployment of people for specific tasks and specific periods;
- a strong set of central operational capacities which are capable of expanding and contracting on the margin, according to need. I will return to the issue of operational capacity later.

While we strongly encourage client-sponsored survey work, some ground rules are enforced:

- no client-sponsored work is carried out on a privileged basis, *i.e.*, with results that are kept private;
- Statistics Canada maintains full professional control, subject only to having to meet the substantive needs of the client. Professional control includes questionnaire content and design, sample design, the collection operation, and processing. It also reserves the right to use the resulting information in its own analytic publications;

- Statistics Canada will not carry out surveys in certain fields which are incompatible with its mandate.
- These include political polling, as well as questions which the public might regard as offensive;
- Finally, Statistics Canada is not in competition with the private sector: it is engaged in large scale surveys with high quality requirements which the agency is in a better position to carry out cost-effectively and for which the agency's stamp of professionalism and legitimacy is important.

6.3 Elements of Professional and Operational Capacity

Maintaining a strong professional and operational capacity is a prerequisite of adaptability. It is particularly important to safeguard it consciously during periods of budget reduction because infrastructure represents a tempting target whose weakening has no immediately visible impact on outputs. Research, analysis, and a methodology capacity are particularly vulnerable.

A strong professional and analytic capacity is particularly necessary for being able to recover from adversity. Indeed, without it we may start a downward spiral of both credibility and resources. We will need our professional staff to develop programs as and when the opportunity presents itself. In addition, they may be able to create informed demand through analytic work which highlights the relevance of statistical information and, whenever appropriate, identifies important gaps in the empirical base needed to support significant conclusions. The analytic capacity of professional staff is also needed for the development or refinement of the conceptual frameworks. In turn, as discussed earlier, such frameworks are prerequisites for the development of relevant new data systems.

Methodology is part of the essential professional capacity of the agency. Our reputation depends on the solidity of our statistical methodology. It might be argued that in times of budget constraints we do not need to have a strong survey design capacity since we are not very likely to launch many new surveys. Yet we have found that improving scientific method can be an important contributor to overall efficiency. This can come about through better survey design and through the development of generalized measurement and processing tools. While it can be destroyed in months, it takes years or even decades to build a strong methodological capacity.

Much the same can be said about the mix of tools and competencies that add up to operational capacity. A well maintained and classified register of businesses; the core supervisory staff of operational entities around which we can build up or reduce operational staff, according to need; the full range of skills needed to maintain a flexible, demand driven informatics capacity; and so on.

It is not enough to simply "preserve" each of these capacity areas. Each must prepare itself for the future by adapting its processes to handle new technology, new methodology or changing respondent attitudes.

Beyond any of the particular elements of operational and professional flexibility, what is needed is particular attention to preserving a spirit of research and innovation at all times. No statistical agency can survive for long without it. Simply carrying on with the same programme, perhaps periodically reduced in response to budget cuts, is a recipe for eventual irrelevance. Yet innovation is particularly difficult to maintain in periods of budget stringency when experimentation has to compete for funds needed to preserve important existing outputs. It therefore requires particular attention. In Statistics Canada we have a planning system (Fellegi 1992) which facilitates this process. Good year or bad, we set aside 2-3 per cent of our budget to support new initiatives. Such a reallocation helps to maintain the intellectual curiosity and ferment that is so characteristic of healthy organizations. A portion of the reallocated funds are used to support pilot surveys and small scale tests which can be used to demonstrate to key client departments how new information could help them anticipate, decide, and monitor policies and programs.

An element of organizational flexibility relates to making multiple exploitations of data as easy as possible. To describe fully the elements of a strategy would take us too far from the theme of the present paper (Fellegi 1991a). Here I would only emphasize the need for three broad approaches. First, we need to create and maintain a single electronic window on all publicly available (*i.e.*, non-confidential) national statistics. This should be the infrastructure supporting all dissemination, from publications to Internet access. Second, behind a publicly available data base of aggregate statistics, we should create and maintain an internal micro data base that encompasses all survey holdings, is fully accessible to all internal staff, and which is well documented. Finally, I would favour all measures designed to place micro data in the public domain – of course, subject to confidentiality. Given the skewness of most of the relevant distributions, we have not found a way to release micro data from most economic surveys. But, after suitable treatment, we release most household and social surveys in this form. This facilitates their use by external researchers, including those in policy departments, as inputs to policy models.

As mentioned before, a major determinant of organizational flexibility is a planning system (Fellegi 1992). In turn, planning must be supported by a detailed project based cost accounting system. These two systems are indispensable to our ability to review regularly the cost structure of our products, to assess the current priority of each product, and to estimate the cost of adding or eliminating particular activities. It would have been exceedingly difficult to manage effectively our response to the last fifteen years of regular budget cuts, punctuated twice by a major injection of funds, without the facility to assess regularly both the substantive and cost implications of modifying our product line.

Perhaps the single most important determinant of organizational flexibility is its human resource strategy (Statistics Canada 1997). Planning decisions invariably imply the reallocation of resources, and people invariably represent the largest single component of project expenditure. We lose the effectiveness of our planning system if we cannot successfully and easily redeploy them according to need. Statistics Canada, like most other statistical agencies, used to be characterized by narrow vertical career paths: if you started work in health, education, labour or manufacturing statistics, it was highly probable that you also ended it in those same fields.

About ten years ago we realized that, for a number of reasons, we simply could not afford to continue along the same path.

- We could not afford the rigid allocation of resources that this implied. On the contrary, it was imperative that we should be able to adjust our programmes in line with client needs and with the available budget, but without the extra concern of having to redeploy people possessing non-transferable skills.
- The regular reduction of budgets substantially reduced the opportunity for advancement, so we had to find an alternative way of maintaining interest and motivation. We found that, for most people, the opportunity and active challenge of new assignments worked well.
- In spite a long series of budget reductions, we wanted the agency to have a degree of robustness enabling it to respond to new challenges. This could only be achieved by acquiring a well trained and flexible staff for which accepting new assignments is a way of life. During the last several years we have developed and implemented a thoroughly integrated training program. We also tripled our training expenditure: from about 1% of total budget to 3%.

Indeed, our organizational robustness is currently subjected to a major test. We have received a substantial injection of new funds to carry out a major expansion of our economic statistics program. The total new funds represent 10 per cent of our total budget but involve as much as doubling the staff in a few divisions. Furthermore, the new statistics are needed for the administration of a high profile government program and, as usual, needed urgently. We simply could not have mobilized the required staff in the short time that was available to us without the preceding ten years of staff rotation and large scale training program.

7. IMPLICATIONS FOR STATISTICAL SYSTEMS: ELEMENTS OF AN EXTERNAL STRATEGY

Just as important as the internal preparedness to meet the challenges ahead, statistical agencies need an "external strategy" as well. Of course, the two sets of strategies interact closely and must be in harmony.

The external strategy should have at least three pillars. The first one, which is so well understood that I will not discuss it here, involves our "core values": maintaining the scientific integrity of the statistical system; safeguarding the confidentiality of identifiable statistical returns; respecting society's privacy norms; and minimizing reporting burden (particularly on small business) through the exploitation of administrative records, sampling, and other statistical techniques.

The other two pillars of the external strategy are relevance and political independence. There is the potential for conflict between these two basic objectives: the closer the statistical system is to the policy process, the higher its potential for relevance – or so it is argued; but such closeness can result in diminished political objectivity, or at least the perception of it. The best solution of this potential conflict depends on national circumstances.

7.1 Achieving a High Level of Relevance

Abstract goals like relevance are achieved through reliance on particular mechanisms. The following are the ones that are most important for Statistics Canada.

(i) Close and Productive Interaction with the Highest Levels of the Bureaucracy

I do not subscribe to the theory that official statistics should aim to satisfy only the needs of the national government, or even those of all levels of government. But I do believe that our priority should be to provide an information base for public policy. The open provision of objective information about public policy issues is of benefit not only to the government, but also to the opposition, to interest groups, indeed to the entire public. It is therefore very important to be well connected with the makers of government policy at the highest level in order to obtain the earliest possible indication of evolving concerns and future government priorities.

The close and productive interaction that is needed does not occur by itself. It evolves over time in response to organizational arrangements and personal initiatives. For example, in Canada the Chief Statistician is a full member of the deputy minister (Permanent Secretary) community and participates in their regular weekly meetings, in periodic retreats designed to "brainstorm" the implications of government priorities, in numerous working groups formed to explore specific issues in depth. Membership in the "club" of deputy ministers opens opportunities to make issue oriented presentations based on statistical information, or to draw attention to important new insights as and when they are released. The primary objective is to ensure that the policy development process takes full advantage of insights that can be derived from statistical information. A not negligible secondary objective is to generate an awareness, at the highest level of the bureaucracy, that statistical information is essential for the policy process and that its usefulness is enhanced, not diminished, by its non-political objectivity.

While there is no substitute to the high level interaction, it is not sufficient. It is essential that, in addition, there should be a web of bilateral interactions (Fellegi 1996) with all major policy departments, as well as those who are guardians of significant administrative record systems of potential statistical interest – such as the customs and taxation authorities.

(ii) Analytic Activities

A strong internal analytic program contributes to an improved understanding of the needs of external analysts – in or out of government. Such an understanding is needed to identify priority data gaps, *i.e.*, information which, if it were available, would make a signal contribution to the understanding of key public policy issues. Such insights are prerequisites of broad support for new initiatives.

Good analysts have a strong personal motivation to explore issues and, more often than not, their explorations result in either data development or in the identification of important gaps. Either way, they will champion the cause of further improvement of information or of a more fruitful conceptual framework.

(iii) A Wide Network of "Listening Posts"

Priority setting is, in the final analysis, subjective. It is all the more important that our assessment be based on broad and balanced information, secured through a variety of formal and informal consultation. Given Canada's federal structure, we have close consultative mechanisms with the provinces in all fields in which they are interested. External expert opinion is received from over a dozen advisory committees, each devoted to a specific subject. Additional views from the business community are sought through marketing efforts by our account executives appointed to work with major clients. Active liaison is maintained by major business organizations and with representatives of the small business sector. At the apex of consultative mechanisms stands the National Statistics Council – a blue ribbon committee of advisors.

(iv) Partnerships with the Academic Community

The academic community, through its analytic activities, can highlight significant insights derived from statistical data bases. It can also be a partner in building conceptual frameworks; call attention to the need for new information products, review plans for new surveys, serve on advisory committees, review analytic products, and so on. As with all other partnerships, keeping it productive involves some effort. We work with them closely to ensure that we can meet their particular needs for access to statistical data bases, we provide opportunities for some of them to spend sabbatical time with us, we co-author papers with them, participate with them in organizing and supporting scientific conferences, and so on.

Collaboration, over time, can make the academic community very highly knowledgeable about the statistical system. In turn, this enables them to be more effective in calling attention to emerging issues and trends. Some academics become natural contacts for the media on issues dealing with their specialization and it is usually helpful to Statistics Canada when they comment on significant new data or analytic releases.

(v) Media Relations

Relevance is determined not only by the potential usefulness of the information produced by a statistical agency, but also by the extent to which the information results in a better understanding of issues. The media have a very influential role to play because it is through their reporting that most people, including many of our elected representatives, acquire statistical information. So frequent and informative media reporting of statistical information is in the public interest. It is also in the particular interest of the statistical office since frequent and objective media references to its products have a positive cumulative impact on the public's appreciation of the agency.

The single most consequential aspect of media relations is to ensure that each statistical release is accompanied by a highly readable analytic summary of what significant new findings and insights it reached. Other measures involving the media might include: free access to all agency releases; the identification in all releases of a competent spokesperson; being proactive in calling attention to errors or data problems; responding in writing to erroneous or misleading articles; availability of senior staff for media interviews; provision of local area detail in releases where this would likely enhance coverage by regional media.

7.2 The Issue of Political Objectivity and its Perception

Public confidence matters because the value of statistics to society directly depends on confidence in their producers. Since few users can actually replicate official statistics, their readiness to use them is ultimately a reflection of their confidence in the professional integrity of statisticians and their ability to carry out their function free of harmful political interference (Fellegi 1991a). The fundamental importance of objectivity has become even more pervasive because of public skepticism about the political process, and the increasing substitution of "objective" statistics for judgement as a means of distributing a diminishing amount of government funds.

I will single out for attention three basic issues and will leave aside such specific and useful techniques as pre-announced release dates for all major series, external review committees, and so on.

(i) Institutional Arrangements

I have discussed elsewhere (Fellegi 1996) the general arguments for and against a centralized statistical system. However, there is no doubt that centralization makes it easier to maintain political independence. First of all, the protection of this independence is a prime responsibility of the head of the agency. The higher his or her standing in the bureaucracy, the more effectively this function can be carried out. This does not primarily derive from power as such, but rather from the fact that public visibility makes an implied threat of resignation of far greater consequence. Since it almost inevitably results in higher standing for the head of the statistical agency, centralization is preferable from the perspective of political objectivity.

Another basic structural issue relates to the formal character of the relationship between the political process and the statistical system. Here, I believe, there is a potentially explicit trade-off between considerations related to relevance versus political objectivity. On the one hand, this paper argued strongly that the status of the chief statistician as a deputy minister (*i.e.*, head of a government department) is of extraordinary importance in maintaining close and productive relationships with other departments -- which, in turn, are key determinants of long run relevance. However, a deputy minister reports to a minister. In the Westminster model of government it is the minister who is responsible to Parliament, not the public servant.

A reporting relationship to a minister can certainly lead to political interference. An alternative arrangement that also preserves the advantages of centralization involves the statistical office becoming explicitly an agent of Parliament, such as the national bank and government audit organization are in many countries. Such an arrangement represents the most secure way of isolating the statistical office from political independence, but raises the risk of increased isolation from the machinery of government, and hence of reduced relevance.

In the case of Statistics Canada, the ministers responsible for the agency have always had a senior portfolio which was their primary policy responsibility. They maintain an arm's length relationship to Statistics Canada on all issues of statistical policy and programs: all questions about technical issues and program priorities are either referred to the Chief Statistician or are answered with reference to the Chief Statistician. This tradition is reinforced and kept alive by the senior public service and the Privy Council Office (the department directly supporting the Prime Minister) who have a clear understanding of the public policy importance of having a credible statistical agency.

Everyone is also well aware that by now there is such a strong employee tradition of political independence in Statistics Canada that the media would find out about any improper attempt to interfere.

Given such favourable circumstances, the regular departmental status of a statistical office confers only advantages. But one might well come to different conclusions in other circumstances, particularly if the most senior levels of the career public service could not be counted on for their strong support of the political independence of the statistical system.

(ii) Budget Control

Budget control is a basic aspect of non-political independence. If the government could target specific statistical programs through the budget process, this could certainly provide an opportunity to target politically embarrassing statistical programs. And even the possibility of such an event could influence behaviour – on both sides.

Statistics Canada experienced repeated budget cuts during the past 12-15 years. However, the agency was allowed both professional and managerial freedom to implement the reductions. Of course, this meant that we had to be prepared to defend our choices. In fact, our management of the process gained us considerable professional and managerial credibility contributing to our subsequent success in obtaining additional funding for some major new initiatives.

Is there a contradiction between budget control and seeking funds for specific initiatives? Not necessarily. The funding was for new activities that we identified, in partnership with others, as having top priority. Furthermore, once received, the funds became part of our regular budget. While we are obviously honour bound to use it for the advertised new programs, this is neither a formal requirement nor is it in perpetuity.

(iii) The Role of Substantive Analysis

An objective and even-handed flow of analytic output contributes significantly to the image of professionalism and political independence that are so essential for statistical offices. Perhaps more than anything else, this helps to differentiate their public image from that of "the government".

Analytic output by Statistics Canada takes a variety of forms, the most visible being what we call our "flagship publications". These provide monthly or quarterly high profile reviews of the economy, of labour market and income analyses, of the analysis of trends in both health and education, and so on. In addition to publishing a wide range of analytic reports, we have an explicit policy that all our statistical releases must be accompanied by a summary of highlights calling attention to significant economic and social developments.

Both objectivity and relevance are important. Objectivity involves exploring all sides of an issue, avoiding policy advocacy, stating assumptions, highlighting major findings whether or not these reflect well on the government. Relevance relates to the choice of topics: they should deal with issues of clear importance – even though some of them might be controversial. Like our other publications, analytic studies must feature readable highlights and they are very widely quoted by the media.

While a regular and visible flow of analytic output can make a very positive contribution to our image, such a program must be particularly well managed. The output must be subject to peer review so as to verify the scientific validity of the analytic approach. But it must also be subject to what we call an "institutional review", to ensure that the analysis is neutral, that it explores issues in an even-handed manner, and that it does not transgress the fine line separating analysis from advocacy.

7.3 International Collaboration

The last element of the external strategy that I want to touch upon is the need to participate in international work. I believe that the international arena is not an optional luxury. It takes at least the following three forms.

(i) The Traditional International Functions

Under this category falls the work well recognized by our profession for over 150 years:

efforts to harmonize concepts and classifications, mutual professional stimulation, and learning from each other. In respect of harmonization, while always important, the need has already increased dramatically and will continue to do so. The requirement arises from a number of sources: transnational corporations, international negotiators, international organizations who set fees and distribute benefits, researchers for whom international comparisons serve as natural benchmarks, and the general public which already has unprecedented ease of access to national data on the Internet.

(ii) Pooling Intellectual Efforts

While the category above encompassed collaboration, it related either to traditional professional interactions, or to work that could only be carried out by and under the aegis of international organizations (such as the development of international classifications and standards). In recent years, stimulated by the persistent conceptual complexity of certain problems (such as the measuring the outputs of the service industries sector), we formed a variety of informal but structured working groups which meet with some regularity and where the national "membership fee" is the contribution of conceptual/developmental work carried out between meetings. Many of these fora turned out to be productive.

(iii) Transnational Dimensions

There is a third category of international work whose dimensions are still fuzzy, but the need for which is clearly discernible. It relates to the looming problem of tracking the economic contributions and transactions of transnational enterprises. No national statistical office can take a proper account of their functioning since they truly operate in a borderless mode. Consider a manufacturer in Canada which, as part of a transnational car enterprise, produces brakes and exports them world wide for use by the same enterprise. This Canadian manufacturer would report export earnings, value added, profits, inventory, capital stock, and so on, all according to the book keeping conventions of the enterprise. In turn, these may well change over time in response to their assessment of the benefits they can derive from differences in national tax laws.

Furthermore, however complicated the problem of tracking might be in respect of goods, it is substantially more so with respect to services, many of which can cross international borders electronically, and go undetected. It is evident that any progress regarding this issue of increasing importance can only be made through the collaboration of national statistical offices in ways and through fora that are yet to be articulated.

8. CONCLUSION

As the millennium is drawing to a close, it is quite fashionable to try to peer into the future: identify emerging trends and provide erudite analyses of their consequences. My experience with similar exercises triggered by other excuses has not been favourable: in retrospect the most important trends turn out to have been different from those that were anticipated. High profile examples of misdiagnoses abound. One of my favourites is the famous statement made by Lincoln Steffens, the American journalist, when he returned from a visit to the Soviet Union in 1919: "I have seen the future; it works..."

Even when we correctly anticipate, our constraints in responding to them are typically quite different from what we might have expected. As a result, I chose to base my analysis on the scenario identified by a group of senior Canadian policy analysts, and so avoided putting forward my own favourites as to what the key policy challenges of the foreseeable future might be.

I could have tried to create my own futuristic scenarios, *e.g.*, about the impacts of involving computer-communications and their impacts on society and the statistical office; I could have speculated about the withering away of nation states – or indeed the opposite (both perspectives were espoused in a recent 75th anniversary issue of Foreign Affairs, by no lesser authorities than Arthur Schlesinger, Peter Drucker, and Anne-Marie Slaughter (Foreign Affairs 1997); I could have tried to discern trends about whether the recent retrenchment of the role of governments is a secular event or only the currently discernible movement of a giant pendulum. I chose to avoid all of that. I fundamentally believe that it is not only possible but essential to pursue a robust strategy that does not depend intimately on our futurology, and I tried to outline the internal and the external elements of such a strategy.

The main feature of the internal capacity is the development and maintenance, even in the face of budget cuts, of a strong professional and operational capacity capable of adapting to its environment. I outlined what I already perceive as major adaptations needed in a number of specific domains of major relevance to public policy. Most of these involve both conceptual and data issues, and I argue that these two issues must be addressed in parallel, in an iterative manner, typically by starting with some conceptual frameworks and then fleshing out and refining them through new data systems.

The key aspects of the external strategy are there to ensure that we have excellent receptors to pick up and filter the signals from our environment; that we place extremely high priority on the various approaches that we need to pursue in order to stay relevant (which means, among other considerations, trying to go beyond our traditional role of monitoring by striving to illuminate issues, including the role of "policy levers"); that we serve our society well – all major groups, and in a manner that is suited to their needs, not our convenience; and last but not least, by doing all that is necessary to preserve and strengthen our non-political and professional independence.

REFERENCES

FELLEGI, I.P. (1991a). Maintaining public confidence in official statistics. *Journal of the Royal Statistical Society*, Series A, 1-22.

FELLEGI, I.P. (1991b). Marketing at Statistics Canada. *Statistical Journal of the United Nations*, ECE, 295-306.

FELLEGI, I.P. (1992). Planning and priority setting – The Canadian experience. In *Statistics in the Democratic Process at the End of the 20th Century*. Anniversary Publication for the 40th Plenary Session of the Conference of European Statisticians, edited by Hölder, Malaguerra, and Vukovich.

FELLEGI, I.P. (1996). Characteristics of an effective statistical system. *International Statistical Review*, 64, 2, 165-198.

FELLEGI, I.P., and WOLFSON, M. (1997). Towards systems of social statistics. *Bulletin of the International Statistical Institute*, 51st Session, Istanbul.

FOREIGN AFFAIRS (1997). 75th Anniversary Issue, September/ October .

PICOT, G. (1997). *Statistics and Public Policy: The Case of Labour Markets and Firms.* Statistics Canada. Paper presented at the meetings of the Statistical Society of Canada, June.

STATISTICS CANADA (1997). Human Resource Development at Statistics Canada. Internal Document, July 21.

THE ECONOMIST (1996). The hitchhiker's guide to cybernomics.' Feature article of September 28.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 129-138
Statistics Canada

Cumulating/Combining Population Surveys

LESLIE KISH[1]

ABSTRACT

Designs for and operations both of multipopulation surveys and of periodic surveys have become more common and important. The needed large resources, both financial and technical, have been organized only in recent decades, and the great values of both became recognized. For both types of designs the developments have concentrated on comparisons between surveys. Yet the coordination and harmonization needed for comparisons also makes the combinations of the survey statistics possible, desirable, and practiced. But the combinations of surveys have been achieved and presented largely without a theoretical/methodological framework, and often poorly. Here such a framework is attempted. Some closely related designs are also discussed: multidomain designs, rolling samples, combining experiments, and combining several distinct survey sites.

KEY WORDS: Multipopulation design; Multidomain design; Periodic surveys; Rolling samples; Combining experiments.

1. INTRODUCTION: MULTIPOPULATION MODELS

A paraphrase of the standard model in all books on survey sampling goes roughly thus: "The aim of survey samples is to produce an estimate of the total Y (or the mean $\bar{Y}$) for a variable Y_i in a population of N elements." Such statements are misleading because they fail to describe the actual purposes and practices of survey sampling. First, most surveys treat many variables, and second, survey results use diverse kinds of statistics; thus sample surveys are "multipurpose" on several dimensions (Kish 1988). But instead of discussing all the omissions of the standard model, I want in this paper to concentrate only on its insufficiency and inadequacy because of its restriction to a single, finite population. Among the several examples below of multi-population expansions that are possible with a new and different model, I begin with two important examples of survey samples that achieve a variety of treatments and results on different dimensions. First is the emergence of multinational designs since 1965, best illustrated by the World Fertility Surveys (section 3), which involve combinations across national spatial boundaries. Second are combinations of periodic samples, best illustrated by "rolling samples" (sections 4 and 5), which concern combinations across temporal dimensions.

The designs and operations for periodic surveys require large resources and new methods. Those for multinational surveys are even more demanding. Both of these types of complex surveys are rather late arrivals among sample surveys and both types are growing in numbers and in importance. Furthermore, both types have been designed and used mostly for comparisons: temporal and spatial comparisons, respectively. The concept of using them additionally for combinations and cumulations is new, and is often encountered initially with doubt and disbelief (sections 3, 4, 5). For both types, the variations between the populations are commonly affirmed as obstacles to combinations or cumulations, and thus are then used for restricting the sample estimates to single populations, because typically methods for combining them are unknown or unavailable. Or even when they are combined, only *ad hoc* methods are used, without justifying them. References to several papers indicate my concern for designs of multinational surveys and of rolling samples. In this paper the emphasis will be on combinations for multinational samples and on cumulations of periodic and rolling samples.

You may notice that I use the terms "cumulating" and "combining" interchangeably and perhaps confusedly. "Combining" seems to fit the multipopulation and multidomain situations better, whereas "cumulating" seems better for periodic and rolling samples. It would be better to have one word to cover both spatial and temporal combinations/cumulations, but neither seems to be exactly right. Also "combinations" serves uses other than joining populations – the usage I wish to emphasize here.

I am also not clear if it is better to consider the enlargement of the scope of samples from one population to several as a new model or as a paradigm shift. Discussions with a few philosophers here left me confused about this choice. And my fellow statisticians probably do not care whether we write the word model or paradigm. In any case, a new model instead of the standard model of sampling from a fixed frame of a stable, finite population is the radical proposal I am pursuing in this paper.

2. MULTIDOMAIN DESIGNS

Statistics for national samples are commonly based on combinations of domains, and these are often quite diverse. But because these combinations are simple and familiar,

[1] Leslie Kish, ISR, University of Michigan, Ann Arbor, MI 48106.

they can also serve as heuristic examples for the less familiar combinations I want to discuss, such as multinational and multiperiodic statistics. The diversity of domains may be recognized within national sample designs; *e.g.*, provinces, which may number from 5 to 20 in most countries. In samples of smaller populations (cities, institutions, firms, *etc.*) similar partitions into major domains also are typical. But for smaller and more numerous domains (*e.g.*, the 3,000 counties of the USA) deliberate sample designs are not feasible for most samples of limited size. For these small domains, methods of "small area estimation" have been developed (Kish 1987, 2.3; Platek, Rao, Särndal and Singh 1987 pp. 267-271). There are great practical differences in both design and estimation between large and small domains, and it is careless to use the adjective "subnational" to cover both. Furthermore, these distinctions between large and small domains exist not only for national designs, but also for samples of smaller populations also. It seems that the structured (nonrandom, grainy) natures of populations persist also on smaller scales. This conforms to the proposed new model of populations, and is supported with empirical analysis of multistage components (Kish 1961).

Although practical for provinces, deliberate designs are not feasible for most domains, whether few and large or many and small, of the kind we call "crossclasses," such as sex and age or occupation, social class, education, *etc.* These "crossclasses" are often important both for their relations (correlations) with the survey variables and for their great diversity. Thus samples of national (or other) populations are mosaics of domains that are diverse and often highly variable; and we must depend on the properties of large probability samples to yield reliable representations of them. In this sense we perceive that all population samples consist of combinations of subpopulations.

Subclasses designate the representations in the entire sample of the domains that compose the whole population. Crossclasses are commonly the most common types of subclasses in survey analysis: partitions of the sample, for which deliberate selection designs are not feasible. For example, occupation and education classes, behavioral and attitudinal categories, and so on. These can be strong explanatory variables for survey analysis; yet we lack the data and resources not only for pre-stratification, but also even for post-stratification methods. From that extreme of lack of controls at one end, we can move to the other extreme of strong controls by separate samples, which can be designed for major provinces.

For example, different methods of sampling can be used in the different provinces. But more common are designs that use different sampling rates; for example, higher sampling rates for small, or for especially important provinces. Sometimes equal sample sizes $n_f = n/H$ are designed for all H provinces in order to obtain (approximately) equal precisions for all provinces, regardless of their sizes. This equal allocation results in sampling fractions n_h/N_h that are inversely proportional to province sizes. But for fixed total sample size n the consequences are higher variances for the entire sample, as well as for crossclasses; see section 8 (Kish 1988). We assume here, that the statistics $\bar{y}_h$ of the provinces (domains) get weighted with population weights $W_h = N_h/\sum N_h$ for the overall statistics $\bar{y}_w = \sum W_h \bar{y}_h$, as is commonly practiced for national statistics. This serves as a useful introduction to the multinational statistics coming next.

3. MULTINATIONAL SAMPLE DESIGNS

National "representative" samples were started by Kiaer (1895) only in 1895 and, after much opposition, they became widespread only after 1945 (Kish 1995). Since then the efforts of the samplers were encouraged and supported by statistical agencies of the United Nations, especially the UN Statistical Office and the FAO. Their spread then led naturally to multinational comparisons of surveys; yet the deliberate design of multinational samples that could provide valid comparisons is recent, starting only around 1965 (Szalai 1972; World Fertility Surveys (WFS) 1984; Kish 1994). The new demands for survey designs for multinational comparisons create many new difficulties: in resources – financial, institutional, cultural; and also in methods. Those difficulties encountered with comparisons reappear also in similar form for multinational combinations, our main concern in this paper.

It is interesting to compare these difficulties with ones with which we are familiar in multidomain designs. From a theoretical perspective, combining the provinces of a country is similar to combining the nations of a continent. Indeed we should profit from those similarities by using metaphorical arguments from the familiar multidomain designs to the proposed multinational combinations. However, from a practical view we find great differences between the two efforts because of five fundamental practical obstacles that make multinational designs much more difficult to achieve, discussed below.

1. The centers of decisions reside in separate national offices, both for setting policy targets and for obtaining funds. Further, within any nation the agencies for policy setting and for resource allocation may be distinct and separate; *e.g.*, the Education Ministry may share participation in a school survey, but the Parliament or the Finance Ministry may fail to allocate funds.

2. The needed technical resources reside in and are staffed and developed by separate national offices. These separate offices may have very different levels and types of technical development, as well as distinct organizational structures and different social connections.

3. The survey variables can vary immensely across national boundaries, due to different cultures, religions, economic and educational levels, legal and social relations, *etc.* Achieving comparable results demands immense efforts – but the task is not impossible, as multinational surveys have shown.

4. The crossnational translation of concepts and of questionnaires, also of codes and analysis, are daunting challenges that need ingenuity, knowledge, and devoted effort.

5. Separate samples must be designed and operated to meet distinct national conditions, with local resources, sampling frames, and field operations. This subject needs volumes; more discussion and study than is possible here.

Multinational comparisons probably go back many years, based on diverse kinds of observations – by travel, wars, conquests, *etc.* But probability sample surveys of entire nations have become common over all continents only during the past half century. As the second phase of development, those national surveys soon led to multinational comparisons. The third phase of deliberate multinational designs dates only from 1965: the Time Use Surveys of 1965 (Szalai 1972); the World Fertility Surveys of 1972-82 (WFS 1984; Cleland and Scott 1987); the Demographic and Health Surveys since 1985 (DHS 1991); the Labour Force Surveys of the European Community (Verma 1992, 1999); see Kish (1994). Other multinational survey designs are also emerging, with the funding and technical resources increasingly meeting the growing effective demands. I am heartened and amazed at the emergence of the International Surveys of Psychiatric Epidemiology, a field that I had feared was beyond the reach of probability surveys in my lifetime! (Heeringa and Liu 1999).

Now, for the new fourth phase, I propose deliberate designs for combinations of multinational surveys. Multinational combinations of surveys are now being produced and published; *e.g.*, European unemployment rates or birth rates; African or Sub-Saharan birth rates or death rates; world growth rates; and many other rates, means, and totals. The data for each nation may be based on probability samples (phase two), or even designed for multinational comparisons (phase three). But the methods used for combining them seem to be completely *ad hoc*; and current usage for the relative weights for combining national statistics seem to be in order of A, B, C, D, E from most common to the least. I made no actual counts, nor an empirical study, but glaring examples appear weekly. Very often the methods and weights for combining the national samples are not even mentioned in the media, even in respectable and scientific journals. To the contrary of the above order, our preferences may be almost in the reverse order of E, F, D, C, B, A – and very much depending on the situation, sample sizes, *etc.*

Allow me, with due modesty, to propose that phase five should be the development of solid theory for choosing among those preferences, and also others. But the need for methods for combinations cannot wait for the future better theory; and it is usual in statistics (and in the sciences) for practice and methods to develop before, and thus both to precede and to stimulate theory. Meanwhile, the discussions below may lead to some improvements in methods, even if they are not quite "optimal."

Here then follow six possible alternative ways and weights for combining national statistics.

A. Do not combine: publish only separate national statistics. This is the most common treatment for several reasons. 1. The authors have not thought of the possibility or need for combination, or rejected them. 2. Perhaps they could not decide on the "best" method, and wanted to leave that to the reader, user, customer. This may be defended by "caveat emptor," or "Bayesian" arguments. However, I reject them. The authors should do no worse in choosing than the average users – who in any case can reject the authors' combination if the national statistics are also published. I believe that when the reader's eye roves over the usual horizontal (or vertical) bars in graphs or over data in tables, it tends to yield a simple mean, hence this roving reduces Method A to Method C in effect. This tendency can perhaps be improved if the width of the bars is made proportional to population weights.

B. Even in the absence of combining populations, designs for multinational comparisons should be "harmonized" in survey measurement methods, to allow for proper comparisons (Kish 1994).

C. Use equal weights ($1/H$) for every country. This method is also common and also avoids (like A) the difficult questions of how to choose population weights W_h, with h denoting country. Probably its use is seldom based on deep reflection, but is widespread mostly because it appears to be a "common sense" approach. Perhaps it would be justified with models, where the between-country variation is paramount, and the population sizes are not relevant. However, I have no faith in such models.

D. Weight with sample sizes n_h. Thus $\bar{y}_w = \sum n_h \bar{y}_h / \sum n_h$, which results automatically from simply cumulating sample cases from separate countries, or sites, or surveys. This is also done frequently, and can be justified when elements are drawn essentially from the "same population" or when per-element variance is the only (or prime) component of variation. It denotes "cumulating cases," as distinguished from combining statistics (Kish 1987, 6.6). This approach can be extended to situations where there are serious differences of element variance due to "design effects"; and then "effective sample sizes" n_h/deff_h may be substituted for the n_h. The "effective sample size" may also be applied if the σ_h^2 differ between populations in order to use weights with precisions $n_h/\sigma_h^2 = 1/(\sigma_h^2/n_h)$. In most situations, however, the variations in sample sizes n_h depend on arbitrary, haphazard factors; and *C* may be a worse choice than using equal weights $1/H$ for all countries (surveys, sites).

E. Use population weights W_h. Thus $\bar{y}_w = \sum N_h \bar{y}_h / \sum N_h$ and $W_h = N_h / \sum N$. This method has the most commonly understood meaning when the N_h represents total numbers of persons in population *h*. However, sometimes the population content may be quite different. For example, for grain (or wine) production it may be total number of farmers, or wheat (or grape) farmers, if those numbers are available, or can be estimated; these populations may yield potentially interesting meanings either for comparisons or for combinations. The population extent also needs to be determined; for example, all persons, or only adults, or only women, or only married women; only urban or rural, or both? Also the timing (date) of the surveys needs standardization, *e.g.*, censuses are conducted in '0 (or '9, or '1) years. Often the population weights are not persons, but acres of land, or tons of steel, or barrels of oil, and so on.

F. Use post-stratification weights. Often in multipopulation situations we encounter the same problem as described later in section 7 for multiple sites. And we may consider the same hierarchy of alternative treatments, the last of which (F) is using "post--stratification" weights. We may well have comparable surveys from several diverse countries of a continent (or the world), but neither all the countries, nor a probability sample of them. (For example, the African or South American countries in the World Fertility Surveys or the Demographic Health Surveys.) One may think of constructing "pseudo-strata" from which the available countries would be posed as "representative selections." Some one stratum could have only a single, available, large country. Another stratum could have 2 (or 3) countries, but with only one available representative that would get the weight of all 2 (or 3) countries. This artificial "pseudo-stratification" procedure may be preferable to simply adding up the available countries into an artificial combination with $W_h(E)$ or with $1/H(C)$. The rationale for this preference is not very different from methods of adjustments for nonresponses.

Several questions and decisions remain concerning the choice among alternative weights. First: the choice should be made chiefly on substantive grounds. What must the combination represent mainly? My own preferences tend strongly toward D and E, and I deplore the prevalence of A and B that we encounter daily. However, I cannot support my preferences on technical grounds. Also I have faced grave problems with the extremes posed by the giants China and India, each more like a continent, and neither solutions E or C seem adequate. I advise defying the geographer's classification of Asia and leave both of them out of Asia, considering them as separate entities. For example, I have omitted all four countries greater than 200 millions in total population (including the USA and USSR) in my computations in 1970 (Kish 1976, Table 4; Kish 1987, 7.3D).

Second: Is the bias due to using incorrect weights important? This would be difficult to prove, as the bias is a function of correlations between the weights and specific survey variables. However, the proof should belong to the denial, as it does with the biases of nonresponses or of poor sampling methods. Ignorance of sources of bias does not imply their absence. I believe that using equal weights instead of population weights can often lead to important biases.

Third: When samples are (roughly) equal-sized, weighting up to population sizes can greatly increase variances. These increases in variances due to unequal weights can be measured quite well (see section 8). They should be balanced against probable biases in models for reducing mean-square errors. In small samples the large variances may dominate the MSE.

Fourth: It seems clear that the combination of population surveys into multipopulation statistics needs a good deal of research, both empirical and theoretical – and especially together.

4. CUMULATING PERIODIC SURVEYS

Periodic surveys have been designed and used mainly for measuring periodic changes, and also for "current" estimates, exploiting the advantages of partial overlaps. But here we shall explore their design and use for cumulated estimates. Furthermore, I include periodic surveys here in order to emphasize their basic similarities to surveys combined over space, such as multidomain and multipopulation surveys. We cannot enter here into the philosophical issues involved in repeated studies of the "same" population, except to note that the "stability" of any population differs greatly for diverse variables (Kish 1987, chapter 6); and the stability for any one variable will also differ greatly, depending on the length of the periods, which may be weekly, monthly, or quarterly. These are common and useful man-made periods. But there exist only two global "natural" cycles of variations: the diurnal and annual cycles, based on the earth's rotation around its own tilted axis, and around the sun.

I must note four practical, rather than theoretical, differences between cumulating periodic surveys and combining multinational or multidomain surveys.

1. Periodic surveys are designed for the "same" population, which tends to retain some stability between periods. The "sameness" and "stability" are only

relative, and with many exceptions; *e.g.*, epidemics in health data or fluctuations in stock prices. They differ greatly between variables and decrease for longer periods.

2. These stabilities imply positive correlations between periods, encouraging designs with "overlapping" sampling units in order to reduce both unit costs and variances for estimates of change and of current values. These overlaps are not desirable for cumulations, so this conflict between the two designs must be resolved.
3. Because similar methods and designs are feasible and generally preferred, they are used over all the periods; on the contrary, harmonization of methods is difficult to achieve between national samples. I emphasize here cumulating periodic surveys, but these aspects also apply to comparisons.
4. Methods for periodic surveys for comparisons have been widely published, in contrast to the novelty both of multinational designs and of periodic cumulations.

There now exist several cumulated representative samples (CRS) of national populations: samples designed for cumulations over large populations. These remain restricted within selections of primary sampling units in order to reduce field costs, whereas "rolling samples" (section 5) are spread deliberately over all sampling units in the population. The Health Household Interview Surveys (HHIS) of the USA are separate weekly samples of about 1,000 households, cumulated yearly to 52,000 households (National Center for Health Statistics 1958, pp. 15-18). These samples are selected by the US Census Bureau within their large sample of PSUs. The yearly samples of over 150,000 persons constitute a remarkable example of multipurpose surveys, representing even rare diseases. The Australian Population Monitors have quarterly nonoverlapping samples that are cumulated to yearly samples, and these are also confined into primary sampling units (Australian Bureau of Statistics (ABS 1993)). The new Labour Force Surveys of the United Kingdom publishes each month the cumulation of three separate, nonoverlapping monthly samples (Caplan, Haworth and Steele 1999). There are other examples as well, and the applications of cumulative representative samples (CRS) are increasing in scope and diversity, although until now they have lacked a common name and literature. Nevertheless, I propose to differentiate the CRS from rolling samples for practical reasons (section 5).

Two problems and methods associated with cumulated samples deserve brief mentions, but with references to more adequate treatments. Asymmetrical Cumulations refer to proposals and some actual practices of reporting large aggregates frequently, but reporting on small domains only after cumulating over longer intervals. For example, the HHIS above may report some national averages each week or monthly, but smaller regions, or specific diseases, only for annual aggregates (Kish 1990).

A serious conflict can arise if periodic samples are to be used (as they should) both for measuring periodic changes and current levels and for measuring cumulations over the periods. This double use has been proposed and practiced, although I do not yet know of any deliberate double designs. Most periodic surveys use partially overlapping samples with some kind of rotation design. One reason often given for these overlaps is the reductions in variances per sample element both for measuring changes between periods and for making current estimates. These reductions depend on positive correlations between the overlapping sampling units. Such reductions are well documented in sampling textbooks and articles since the original papers on this topic (Jessen 1942; Patterson 1954). But even greater reductions are possible in element costs, when the later interviews are much cheaper than the first contacts; for example, if the later contacts are by telephone. On the other hand, separate new samples will be much preferred for cumulations in order to avoid the positive correlations. One may imagine different compromises that may be efficient, when: (a) most of the positive correlations are not high; (b) reinterview costs are not much cheaper; and (c) reinterview response rates are discouraging.

However, consider also a new design that I call a Split Panel Design (SPD) that adds a panel *p* to a parallel series of nonoverlapping samples *a-b-c-d etc.*; with the combination then denoted as *pa-pb-pc-pd etc*. The panel replaces the overlaps of rotating designs and provides the useful correlations for measuring net (macro) changes. Further, it also serves to measure individual (gross) changes, which are lacking in the usual designs of overlapping sampling units, because of the mobility of persons and households. Including panels of individuals (persons, elements) would bring considerable advantages for SPD over all current overlapping samples, which usually use merely the same sampling units (Kish 1987, 6.5; Kish 1990).

Another considerable advantage of SPD is that these overlaps would be based on the correlations from all periods, rather than only for the arbitrarily chosen periods for the rotation designs. How arbitrary are these? Some decisions use one-month groups, some three months, others 12 months, *etc*., *etc*. It is most unlikely that these disparate overlaps are actually "optimal" for those countries. It seems most likely that the "optimal" overlap cannot be predetermined for any single variable, and a single optimal period is even less likely for multipurpose designs.

5. ROLLING SAMPLES AND CENSUSES

These should be considered as special types of the related cumulative representative samples (CRS); but rolling samples (RS) should be distinguished, because they are designed for different and specific functions. CRS have been confined to designs of PSUs. They are spatially restricted for cost reasons and for fitting the designs of labor force surveys, and other surveys associated with them. However, RS designs must aim at a much greater spread in order to facilitate maximal spatial range for cumulations over time. Rolling samples must be designed specifically to readily yield good estimates for all small spatial units, when the periodic samples are cumulated into annual or decennial larger samples or censuses.

First let us define a rolling census: it consists of a combined (joint) design of F separate (nonoverlapping) periodic samples, each a probability sample with fraction $f = 1/F$ of the entire population, and so designed that the cumulation of the F periods yields a detailed census of the whole population with $f' = F/F = 1$. Intermediate cumulations of $k<F$ periods should yield rolling samples with $f' = k/F$ and with details intermediate between 1 and F periods.

Imagine a weekly national sample, each designed with epsem selection rates of $f = 1/520$. The cumulations of 52 such weekly samples would yield an annual sample of 52/520=10 percent. Then ten of these annual samples would yield a census of 520/520. I have proposed in several papers to have these rolling samples replace both kinds of the most important forms of official statistics that are either used or planned in many countries: the monthly surveys of population and labor force and the decennial censuses. Even more important, these surveys could also provide annual detailed data, perhaps with 10 percent samples, which are badly lacking, and needed in many countries (Kish 1990, 1997, 1998). Providing spatially detailed annual statistics for a variety of economic and social variables, not a mere population count of persons, would be the chief aim of rolling samples in many countries. These are needed even in countries that can provide fairly good estimates of population counts and a few simple statistics either from registers or with estimation methods. In countries without good frequent (monthly or quarterly) surveys of labor force and population, rolling samples could also serve them as efficient vehicles.

I must admit that the above basic ideas provide merely the skeleton for any actual national design for rolling samples. But such actual national samples have been recently designed – the largest and best of which is the American Community Survey (ACS) – now undergoing a 37-area pilot study by the US Census Bureau (Alexander 1999). This aims to provide monthly surveys of 250,000 households and detailed annual statistics based on 3,000,000 households, after year 2003; and also to provide quinquennial and decennial census samples later. The National Statistical Office of France is working on plans for a Census Continué (Isnard 1999). The Labour Force Surveys of the United Kingdom are now based on cumulated monthly surveys. Some other countries are examining different but generally similar possibilities.

It is also proper to add references to two early publications describing "rolling samples" of large sizes, although not national in scope (Mooney 1956; Kish *et al.* 1961). Others probably exist that I have not seen.

How to cumulate periodic surveys? This topic must receive serious technical consideration in the future, because so far they have been done only with *ad hoc* procedures. Perhaps for cumulating over a single year, epsem samples with the same sampling fraction f, and simple cumulation of cases may serve as a simple model: averaging over seasonal and random variations may outweigh secular trends. However, averaging annual statistics over 10 years may have to consider secular trends in population size.

Consider several alternative sets of weights W_i to be assigned to yearly means $\bar{y}_i$ for a decennial mean $\bar{y}_i = \sum W_i \bar{y}_i \, (i = 0,1,2,...,9)$ and $\sum W_i = 1$.

a) $\bar{y}_{ta} = \bar{y}_9$, with $W_9 = 1$ and the other nine $W_i = 0$, utilizing only the final year. This could be used for national and large domain estimates, and for highly fluctuating variables (unemployment, epidemics, stock prices), where the need for timeliness dominates sampling precision.

b) $\bar{y}_{tb} = \sum W_i \bar{y}_i$, with all ten $W_i = 0.1$. For variables without time trends, and for small domains, obtaining a stable average over time may be good strategy.

c) $\bar{y}_{tc} = \sum W_i \bar{y}_i$, with $W_0 \le W_1 \le W_2 \le ... \le W_9$, monotonically increasing (or nondecreasing) W_i. The curve of increase may be determined with a model or with empirical data. Thus $\bar{y}_{ta}$ and $\bar{y}_{tb}$ may be viewed as two extremes of $\bar{y}_{tc}$. They all seem better than the present practice of giving full weight $W_0 = 1$ to a decennial census that may be from 1 to 10 years old and obsolete.

Furthermore, with rolling censuses, the statistical office need not wait to publish only decennially. It can publish annually the results of the latest rolling samples, with several available alternatives from those above: either the latest year $\bar{y}_{ta}$; or $\bar{y}_{tc}$ an average that favors the latest years. Or "asymmetrical cumulations" favoring $\bar{y}_{tb}$ for smaller domains, but $\bar{y}_{ta}$ for larger domains and totals. It could conceivably publish both $\bar{y}_{ta}$ and $\bar{y}_{tb}$ and let the reader choose (perhaps publish electronically). Clearly technical research will be needed to search for "optimal" solutions to support the applications already appearing.

6. COMBINING EXPERIMENTS

A) This topic has been the subject of three early and good papers by Cochran and has also received attention from both Yates and Fisher at Rothamsted (Cochran 1937 and 1954; Yates and Cochran 1938). These dealt with experiments relating crop yields (predictands) to fertilizers (one or more predictors), conducted over different populations, fields, and years. They used ANOVA methods for statistical analyses and for combining the several independent experiments.

B) Fisher's test for combined probabilities, from 2×2 Chi-square tests of the "same" null hypothesis is even older. It can use entirely different populations, and even diverse variables, for testing the "same" null hypothesis. This well-known test can be found in most statistics textbooks.

C) Methods of meta analysis are newer, and increasingly used. They combine experimental results from different samples and populations for the same predictand (outcome) variable from one or more predictors (inputs). (Glass 1976, Hodges and Olkin 1985.)

Methods for combining sample surveys are just emerging, much later than methods for combining experiments. The two fields, however, have many similar aims, which should be noticed, in order to see useful relations between the two distinct topics. Perhaps these relations can be best perceived by looking at the differences between the aims and the problems that have been the subjects of the two methods. There seem to be three main differences between the two methods, as they have been applied.

1. Combining surveys (CS) needs a great deal of advance preparation, planning, and coordination. This is true of multinational surveys for both the comparisons, which have been already achieved, and for their combinations, which are new. For national multidomain surveys the coordination comes naturally, but for multinational surveys the coordination of the separate national designs is difficult, but necessary (Kish 1994). On the contrary, a great virtue of combined experiments (CX) is that they can be performed on the reports of experiments already performed, as the name meta analysis signifies. That analysis is based on the relations of the predictand/predictor pair of experimental variables. The Fisher test needs only the probabilities P_i achieved by the tests of significance.

2. The second difference between the two methods is related to the first. The CX are based on experiments, whereas CS concentrates on surveys. Thus CX emphasizes experimental control through randomization of variables over subjects. However, CS are based on probability sampling with randomized selections of subjects – not variables – from defined populations. Usually these two kinds of randomizations are difficult to achieve in any research study and one must be sacrificed (Kish 1987, 1.1). The population base of CS is specified, whereas those for CX usually are not and cannot be.

3. Third, CS involves a full statistical analysis, and even a full survey method, designed for similarity and comparability in order to facilitate the joint analysis. On the contrary, the methods of CX can use the very end of the statistical analyses, often even from published statistics. The extreme of this kind of abstraction is shown by the combined Fisher test, based only on the terminal P_i values of the separate statistical tests.

Because of the large, consistent and interrelated differences between Combined Experiments and Combined Surveys, it may be best to keep the two methods separate. Some may propose that the gap between the two subjects is only an historical accident and that the gap can be closed sometimes. But I believe that it is more useful to maintain the separation of the two methods, even if sometimes a compromise may be usefully adopted.

That still leaves open the question whether the three methods of combined experiments (A, B, and C above) should be called "Combined Experiments," as Cochran, Yates and Fisher called them since the 1930s or if it is better to distinguish them all as "Meta-Analysis," now a widely known and accepted joint designation. Happily we need not decide here, but perhaps meta-analysis is the best, provided we also recognize the earlier successes.

7. COMBINING SEPARATE SITES

Suppose that similar data have been collected in several sites of a combined population, but not in all of the sites, nor in a probability selection of them. The sites may be cities, provinces, or districts of one country. Or they may be institutions, such as schools, or hospitals, or factories. Or the sites may even be entire countries of a continent. I have seen a variety of such situations when the sites are either chosen arbitrarily, or are simply "volunteers." Often the sample sizes per site are similar, though the population sizes of the sites vary greatly. Here follows a list of possible alternative treatments of the data.

A. Separate survey estimates $\bar{y}_i$ may be presented only. Usually this is all that is done, especially if the data have not been coordinated, or "harmonized." Any comparisons and any combinations of the separate statistics are left to the readers, to use their own methods or resources.

B. Comparisons between the separate sites require harmonization (of variables, measurements, timing, populations) to render the differences $(\bar{y}_i - \bar{y}_j)$ meaningful.

C. Simple cumulations $\bar{y}_t = \sum y_i / \sum n_i$ of all sample cases amount to assuming that the populations N_i of the sites can be considered parts of the same population of $\sum N_i$ elements. Note that the sample means $\bar{y}_i$ are weighted by the sample sizes n_i. Often these are nearly equal and then C approaches D.

D. Equal combination $\sum \bar{y}_i / k$ of k sites weight each of the sites equally, disregarding both the sample sizes n_i and the population sizes N_i.

E. Weighted combinations $\bar{y}_w = \sum W_i \bar{y}_i / \sum W_i$ weight the sites with some measure of their relative importance.

Population sizes N_i seem reasonable, but others may be used. However, we may object to the combination of an arbitrarily selected set of sites.

F. Post-stratification weights $W_i \propto \sum_j N_{ij}$ can save attempts to overcome the above objections by constructing pseudostrata $\sum_j N_{ij}$, composed of "similar" sites, from which the unit N_i may be considered a valid selection. Thus the total sample then is considered a sample from the larger population of total size $\sum_i \sum_j N_{ij}$. Such model building resembles the attempts to reduce nonresponse bias with nonresponse classes.

Three sets of decisions must be made, and this order is chronological in activity, but not necessarily in planning. a) The allocation of sample sizes, especially whether equal sizes for the sites, or proportional to relative population sizes (W_i). b) Whether the samples should be combined, or to merely accept alternative a). c) What weighting to use among alternatives b) to f).

The above alternatives resemble those in section 3 and multinational combinations may be viewed as special cases of multi-site combinations, but a very special case, for the reasons given there. Furthermore, the alternatives listed above deal not with academic or idle speculation, but with many practical, actual problems. I have advised and argued on problems of every kind, and felt the need for and lack of dependable references on combinations and cumulations, whether technical and published or oral and authoritative. Some examples I have encountered:

a) The World Fertility Surveys had national sample sizes without much (any) relation to population sizes. Should they be combined and how? I thought yes and with $N_i(E)$.

b) Samples of several hundred households were selected in each of 12 large cities of the USA (which had "racial riots" in 1968). Should they be combined and how? I thought yes and with $N_i(E)$.

c) In each of 13 counties of the USA samples of a few hundred 4-year-old children were selected for a study of preprimary learning situations. They were combined with method *F*.

d) In 11 of China's 30 provinces probability samples averaging 1,000 4-year-old children were selected for studies of preprimary learning situations. They were combined with method *F*.

e) In 5 of Nigeria's 30 states small urban and rural samples were selected for studies of preprimary learning situations of 4 year olds. After examining the 5 × 2 small samples the sample cases were merged with Method *C* into urban and rural samples.

f) Coordinated survey designs and university resources are being planned for 5 to 8 large cities of China. The designs are planned both for comparisons and for combination, with either Method *E* or *F*.

8. ERRORS, LOSSES, COMPROMISES

The Mean Square Error of a weighted combination of means may be written as

$$\begin{aligned}&\mathrm{MSE}\left(\sum W_i\bar{y}_i\right)\\&\quad=\mathrm{Bias}^2\left(\sum W_i\bar{y}_i\right)+\mathrm{Var}\left(\sum W_i\bar{y}_i\right)\\&\quad=\left\{\sum W_i\left[E(\bar{y}_i)-\bar{Y}_i\right]\right\}^2+\left(\sum W_i^2D_i^2S_i^2/n_i\right).\end{aligned}$$

This holds for distinct countries (i) and distinct domains like provinces. But for some domains there may also exist covariances (S_{ij}), positive or negative. The relative weights are W_i, and S_i^2 and n_i are element variances and sizes, with design effects D_i^2 to compensate for the effects of complex designs. On any study all these values can differ greatly between variables. Note that the bias of the combined mean is the weighted average of the individual biases. For periodic samples these may be fairly constant. For multipopulation and multidomain samples this emphasizes the need for reducing biases for the larger units, with large W_i. The variances of means decrease in proportion to the number of units being averaged, and thus they decrease in importance relative to the biases.

The situation is different for comparisons, where

$$\begin{aligned}&\mathrm{MSE}(\bar{x}-\bar{y})\\&\quad=\mathrm{Bias}^2(\bar{x}-\bar{y})+\mathrm{Var}(\bar{x}-\bar{y})\\&\quad=\left[E(\bar{x}-\bar{y})-(\bar{X}-\bar{Y})\right]^2+\mathrm{Var}(\bar{x})+\mathrm{Var}(\bar{y})\\&\quad=\left[\{E(\bar{x})-\bar{X}\}-\{E(\bar{y})-\bar{X}\}\right]^2+D_x^2S_x^2/n_x+D_y^2S_y^2/n_y.\end{aligned}$$

Note that the biases of differences tend to vanish if the biases are similar, even when not small. The variance is the sum of two variances (and a small n_x or n_y can increase it), hence may dominate the bias term. When there are overlaps (in periodic surveys) the covariance term $-2\,\mathrm{Cov}(\bar{x},\bar{y}) = -2D_{xy}S_{xy}n_c/n_x n_y$ tends to decrease the variance.

I have emphasized in some detail elsewhere the need for the utmost "harmonization," for the coordination of survey methods: in variables, measurements, and in populations. On the other hand, there is great freedom to choose different sampling methods for the different populations, provided they are all based on good probability samples (Kish 1994).

In multipopulation combinations, frequent and serious conflicts arise, because the relative sizes W_i of the populations (of countries or of provinces) often vary greatly;

ranges of 1 to 50 or more are common. But the sample sizes may be (roughly) equal for all H populations. Then weights k_i may be introduced to adjust the combinations to the W_i. These inequalities of sampling rates increase the variances of combinations by a relative factor $1 + L = 1 + C_k^2$; where L denotes relative loss (increase in variances) and C_k^2 the coefficient of variation among the weights k_i. Both are zero when all k_i are the same, *i.e.*, for proportional allocation of the n_i to the W_i. But then the average variance of the populations and their comparisons suffer even greater losses than the sum. This conflict can be resolved with compromises, especially an "optimal" compromise with $n_i \propto \sqrt{(W_i^2 + 1/H^2)}$ (Kish 1976, Kish 1994).

A good numerical example comes from the 10 provinces of Canada, whose total population (in 1991) of 27, 211,000 with an epsem selection of $f = 1{:}2721$ would yield roughly these 10 values of n_i in row 1 for a total of $n = 10{,}000$. You see that the largest province of 3,706 cases is about 75 times greater than the smallest with 49. This range seems common for provinces within most countries. Also for multipopulation cases; *e.g.*, in the European Union, Germany is about 200 times the size of Luxembourg. The proportions are $W_i = n_i / 10{,}000$; and a proportional sample would yield an optimal value for $\sum W_i \bar{y}_i$ of $\sum W_i^2 \bar{y}_i^2 / n_i$, hence a relative loss function $1 + L = 1$, with loss $L = 0$. For simplicity and to concentrate on weights, we can assume that element variances $D_i^2 S_i^2$ and costs c_i are similar, or can be averaged out. However, these proportional n_i values would result for average provincial means $\sum \bar{y}_i / 10$ or for average comparisons of provincial values $(\bar{y}_i - \bar{y}_j)$ of $1 + L = H \sum (1/W_i) = 3.9785$ for a relative loss of 2.9785, a 300% increase in average variances. These losses come mostly from the 6 small provinces (Derivations in Kish 1976).

Row 1	3,706	2,534	1,206	935	401	363	331	266	209	49
Row 2	2,437	1,730	995	869	684	676	669	657	648	636

Thus, some people (in Canada and in other countries too) ask for equal size samples, $n_i = 1{,}000$, so that each province can provide the same precision. Then the means $\sum \bar{y}_i / 1{,}000$ will all have variances $\sum (1/1{,}000)$ and relative efficiency of $1 + L = 1$, with loss $L = 0$. However, the national mean will have a variance of $\sum W_i^2 / 1{,}000$, with a relative loss of $1 + C_k^2 = H \sum W_i^2 = 2.3003$, or a 130% increase in variance. We must also remember that all crossclasses, such as those by age, education, occupation, *etc.*, will also tend to suffer similar losses.

However, some remarkably good compromises can be had, and the best is a least-square solution with the $n_i \propto \sqrt{(W_i^2 + H^{-2})}$. These give the $1 + L$ values of $1 + L = 1.2424$ and $1 + L = 1.2630$, for $\sum W_i \bar{y}_i$ and $\sum \bar{y}_i / H$, respectively, only a 25% loss for each! The n_i values in Row 2 show a "floor" between 600 and 700 for the n_i for the 6 small provinces, and a roughly proportionate increase (but below 10,000 W_i) for the largest 4 provinces. This optimal allocation has in fact been used for some of the surveys of Statistics Canada (Tambay and Catlin 1995). It is interesting that the mathematical solution also makes good common sense (Kish 1976, 7.6, Kish 1987, 7.3, Kish 1988). However, the mere common senses solutions of allocations proportional to $\sqrt{W_i}$ are less efficient than the optimal allocation.

REFERENCES

ALEXANDER, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.

AUSTRALIAN BUREAU OF STATISTICS (1993). *The Australian Population Monitor*. Canberra: ABS.

CAPLAN, D., HAWORTH, M., and STEEL D. (1999). UK labour market statistics: Combining continuous survey data into monthly reports. *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.

CLELAND, J., and SCOTT, C. (1987). *The World Fertility Survey*. Oxford: The Oxford University Press.

COCHRAN, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, Series B, 4, 102-18.

COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-29.

GLASS, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3-8.

HEDGES, L.V., and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

HEERINGA, S.G., and LIU, J. (1999). Complex sample design effects and inference for mental health survey data. *International Journal of Methods in Psychiatric Research* 7, 56-65.

ISNARD, M. (1999). *Alternatives to Traditional Census Taking: The French Experience*. Paris: INSEE.

KIAER, A.N. (1895). *The Representative Method of Statistical Surveys*. English translation 1976, Oslo: Statistik Centralbyro.

KISH, L. (1961). A measurement of homogeneity in areal units. *Bulletin of the International Statistical Institute*. 33rd session, 4, 201-209.

KISH, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, Series A, 139, 80-95.

KISH, L. (1987). *Statistical Research Design*. New York: John Wiley & Sons.

KISH, L. (1988). Multipurpose sample designs. *Survey Methodology*, 14, 19-32.

KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 1, 63-71.

KISH, L. (1994). Multipopulation survey designs. *International Statistical Review*, 62, 167-186.

KISH, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55-77.

KISH, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, 14, 31-46.

KISH, L., LOVEJOY, W., and RACKOW, P. (1961). A multi-state probability sample for traffic surveys. *Proceedings of the Social Statistics Session, American Statistical Association*, 227-230.

MOONEY, H.W. (1956). *Methodology in Two California Health Surveys, San Jose (1952) and Statewide (1954-55).* U.S. Public Health Monograph No. 70.

NATIONAL CENTER FOR HEALTH STATISTICS (1958). Statistical designs of the Health Household Interview Surveys. *Public Health Series,* 584-A2.

PLATEK, R., RAO, J.N.K., SÄRNDAL, C.E., and SINGH, M.P. (Eds) (1989). *Small Area Statistics.* New York: John Wiley & Sons.

SZALAI, A. (1972). *The Use of Time.* The Hague: Mouton.

TAMBAY, J.-L., and CATLIN, G. (1995). Sample design of the National Population Health Survey, *Health Reports,* Catalogue No. 82-003, 7, 29-38.

VERMA, V. (1992). Household surveys in Europe: Some issues in comparative methodologies. In *Seminar: International Comparisons of Survey Methodologies.* Athens.

VERMA, V. (1999). Combining national surveys for the European Union. *Proceedings of the 52nd Session of the International Statistical Institute,* Helsinki.

WORLD FERTILITY SURVEYS (1984). *Major Findings and Implications.* The Hague: International Statistical Institute.

YATES, F., and COCHRAN, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Science,* 28, 556-80.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 139-149
Statistics Canada

Managing Data Quality in a Statistical Agency

GORDON BRACKSTONE[1]

ABSTRACT

Confidence in the quality of the information it produces is a survival issue for a statistical agency. If its information becomes suspect, the credibility of the agency is called into question and its reputation as an independent, objective source of trustworthy information is undermined. Therefore attention to quality is a central preoccupation for the management of a National Statistical Office. But quality is not an easily defined concept, and has become an over-used term in recent years. Quality is defined here to embrace those aspects of statistical outputs that reflect their fitness for use by clients. We identify six dimensions of quality: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. For each dimension of quality, we consider what processes must be in place to manage it and how performance can be assessed. Finally, we try to integrate conclusions across the six dimensions of quality to identify the corporate systems necessary to provide a comprehensive approach to managing quality in a National Statistical Office.

KEY WORDS: Quality; Official Statistics; Relevance; Accuracy; Timeliness.

1. INTRODUCTION

Confidence in the quality of the information it produces is a survival issue for a statistical agency. If its information becomes suspect, the credibility of the agency is called into question and its reputation as an independent, objective source of trustworthy information is undermined. With this comes the risk that public policy debates become arguments about who has the right set of numbers rather than discussions of the pros and cons of alternative policy options.

Therefore attention to quality is a central preoccupation for the management of a National Statistical Office (we will use the abbreviation NSO to refer to a generic government statistical agency that may go under different names in different countries). Current recognition of the importance of quality to NSO management is reflected in several recent events in the realm of official statistics. For example, *Quality Work and Quality Assurance within Statistics* was chosen as the theme for the May 1998 meeting of the heads of NSO's in the European Community (EUROSTAT 1998); several principles stressing the importance of relevance, professionalism and openness were included among the ten *Fundamental Principles of Official Statistics* approved by the U.N. (United Nations 1994); Performance Indicators (which includes quality as a critical dimension of performance) was chosen as the subject for substantive discussion at the 1999 Conference of European Statisticians (UNECE 1999). This journal through its 25 year history has carried many articles addressing quality issues and it is appropriate that in this anniversary issue we address the topic of quality management in statistical agencies.

But *quality* is not an easily defined concept, so the first issue is what do we mean exactly by quality in this context. Quality has become an over-used term during the past two decades. The Total Quality Management (TQM) movement and other management frameworks have broadened the concept of quality beyond the traditional statistician's concepts of data quality as defined, for example, by the mean square error of an estimator. So our first challenge is to circumscribe the concept of quality as it relates to the work of a NSO. That is the object of section 2 of this paper in which we will suggest six dimensions of quality about which NSO's need to be concerned. In the subsequent six sections we address each of these dimensions in turn, and consider for each: what exactly needs to be managed, what approaches might be used for managing it, and how might we measure performance in managing it.

In section 9 we will attempt to integrate some of the conclusions across the six dimensions of quality, and to identify the agency-wide systems necessary to provide a corporate approach to the management of quality. In the final section we suggest some areas requiring further attention in order to manage quality more effectively.

2. DEFINITION OF DATA QUALITY

The difficulty for statisticians in defining quality as it applies to statistical information is that they thought that was something they had already done. Their whole training is concerned with optimizing the quality of statistical estimates, the fit of statistical models, or the quality of decision-making in the face of uncertainty. Using concepts such as standard error, bias, goodness of fit, and error in hypothesis testing, they have built up methodology for estimation and analysis in which the quality of data, as defined in a certain precise sense, plays a central role.

But the term quality has come to take on a broader meaning in the management of organizations. The TQM movement and other management philosophies have focused on the fitness of final products and services for

[1] Gordon Brackstone, Informatics and Methodology Field, Statistics Canada, Ottawa, Ontario, K1A 0T6, e-mail: bracgor@statcan.ca.

users, have emphasized the need to build quality into the production and delivery processes of the organization, and have stressed the importance of employee involvement in process redesign and commitment to improvement of the final product or service. Statistical methods play an important role in these management approaches, but they are part of a larger picture. A question to consider is how this broader notion of quality applies to an organization engaged in the production and delivery of statistical information. The definition and management of quality in government statistics were discussed in several papers presented at the 1995 International Conference on Survey Measurement and Process Quality (Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewin 1997, de Leeuw and Collins 1997, Dippo 1997, Morganstein and Marker 1997, Colledge and March 1997) and more recently in Collins and Sykes (1999). For an earlier approach see Hansen, Hurwitz and Pritzker (1967).

If we accept that the needs of clients or users should be the primary factor in defining the activities, and assessing the success, of a NSO, we can define the concept of quality as embracing those aspects of the statistical outputs of a NSO that reflect their fitness for use by clients. But, since a NSO has many and varied clients, and each may make a variety of uses of statistical information, this does not provide an operational definition. However, it does allow a more systematic consideration of the most important dimensions of this broader concept of quality, a concept which clearly extends beyond the statistician's traditional preoccupation with accuracy, the aspect of quality which most easily lends itself to rigorous mathematical development.

The first aspect is whether the NSO is producing information on the right topics, and utilizing the appropriate concepts for measurement within these topics. Does it have information relevant to topical policy issues or is it still counting buggy whips? Does it utilize a definition of family that is pertinent to today's society? Does its classification of occupations reflect the current labour market? These are examples of questions about the **relevance** of statistical information.

Given that the NSO is measuring relevant topics using appropriate concepts, is it measuring them with sufficient accuracy? Exact measurement is often prohibitively expensive, and sometimes impossible, so the issue is whether an acceptable "margin of error" has been achieved. This is the traditional domain of statisticians with their concepts of standard error, bias, confidence intervals, and so on. We will refer to this dimension of quality as **accuracy**.

The next two dimensions of quality relate to when and how statistical information is made available to clients. Accurate information on relevant topics won't be useful to clients if it arrives after they have to make their decisions. So the **timeliness** of statistical information is another important dimension of its fitness for use. Timeliness to the day may be crucial for key monthly economic series, but less important for measures of slowly changing phenomena.

For statistical information to be useful, clients have to be able to determine what is available and how they could obtain it. It then has to be available to potential clients in a form that they can use and afford. Both searching facilities and statistical products themselves have to use technology that is available to potential clients. This collection of considerations will be referred to as **accessibility**.

To make appropriate use of statistical information from the NSO clients have to know what they have and to understand the properties of the information. That requires the NSO to provide descriptions of the underlying concepts, variables and classifications that have been used, the methods of collection, processing and estimation used in producing the information, and its own assessment of the accuracy of the information. We will refer to this property of statistical information as its **interpretability**.

Finally, as an extension of interpretability, clients are sometimes faced with utilizing different sets of statistical information derived from different sources and at different times. Appropriate use is facilitated if information can be validly compared with other related data sets. This facility is achieved through the use of common, or at least comparable, concepts and methodologies, across products and across occasions. The degree to which statistical information fits into broad frameworks and uses standard concepts, variables, classifications and methods will be referred to as its **coherence**.

These six dimensions are summarized in Table 1. Clearly they are not independent of each other. For example, all of the other five have an impact on relevance. Accuracy and timeliness often have to be traded off against each other. Coherence and relevance can sometimes be in conflict as the needs of current relevance and historical consistency compete. Information provided to ensure information is interpretable will also serve to define its coherence. Despite these interactions, these six dimensions provide a useful basis for examining how quality in this broad sense should be managed within a NSO.

It is worth noting that most of the important properties of statistical information are not apparent to users without the provision of supplementary information (or metadata) by the NSO. The accuracy of information cannot be deduced just by looking at the numbers alone – some comparisons to other sources may shed light, but the NSO, which alone has access to the underlying microdata and first-hand knowledge of the methodology used, has to provide measures of accuracy. The relevance of information may not be apparent without information on the underlying concepts, classifications and methods used. Only timeliness and accessibility are directly observable by users.

It is also worth noting that relevance, accessibility, and coherence usually have to be considered across a whole set of outputs of a NSO, rather than for each output individually. The relevance of statistical information depends on what else is available and therefore needs assessment across

Table 1
The Six Dimensions of Data Quality

Relevance	The ***relevance*** of statistical information reflects the degree to which it meets the real needs of clients. It is concerned with whether the available information sheds light on the issues of most importance to users. Assessing relevance is a subjective matter dependent upon the varying needs of users. The NSO's challenge is to weigh and balance the conflicting needs of different users to produce a program that goes as far as possible in satisfying the most important needs and users within given resource constraints.
Accuracy	The ***accuracy*** of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure. It is usually characterized in terms of error in statistical estimates and is traditionally decomposed into bias (systematic error) and variance (random error) components. It may also be described in terms of the major sources of error that potentially cause inaccuracy (*e.g.*, coverage, sampling, nonresponse, response).
Timeliness	The ***timeliness*** of statistical information refers to the delay between the reference point (or the end of the reference period) to which the information pertains, and the date on which the information becomes available. It is typically involved in a trade-off against *accuracy*. The *timeliness* of information will influence its *relevance*.
Accessibility	The ***accessibility*** of statistical information refers to the ease with which it can be obtained from the NSO. This includes the ease with which the existence of information can be ascertained, as well as the suitability of the form or medium through which the information can be accessed. The cost of the information may also be an aspect of *accessibility* for some users.
Interpretability	The ***interpretabilty*** of statistical information reflects the availability of the supplementary information and metadata necessary to interpret and utilize it appropriately. This information normally covers the underlying concepts, variables and classifications used, the methodology of collection, and indications of the accuracy of the statistical information.
Coherence	The ***coherence*** of statistical information reflects the degree to which it can be successfully brought together with other statistical information within a broad analytic framework and over time. The use of standard concepts, classifications and target populations promotes coherence, as does the use of common methodology across surveys. *Coherence* does not necessarily imply full numerical consistency.

a whole program. By definition, the same is true of coherence. Most statistical products are delivered through a common dissemination system for the whole NSO so that questions of accessibility are largely corporate too. On the other hand, accuracy, timeliness, and interpretability can be considered as properties of each statistical output, even though, here too, each output may make use of tools or approaches that are common across programs.

We will next consider the management of quality within each of these dimensions.

3. RELEVANCE

Maintaining relevance requires keeping in touch with the full array of current and potential information users, not only to monitor their current needs but also to anticipate their future needs. Information needs are rarely formulated clearly in statistical terms. A major challenge is to translate expressions of interest in particular topics into likely information needs in the future. The relevance of a data set depends on what other data sets are available in related areas of interest. Relevance is therefore more meaningfully managed and assessed at the level of a "statistical program" rather than for an individual data set.

To assure relevance three primary processes need to be in place: client liaison; program review; and priority determination. These are described in the next three sections, followed in section 3.4 by a brief discussion of how performance in the domain of relevance might be assessed.

3.1 Monitoring Client Needs

The NSO requires a set of mechanisms whereby it stays abreast of the current and future information needs of its main user communities. These mechanisms need to include an array of consultative and intelligence-gathering processes to keep the NSO tuned in to the issues and challenges being faced by major users and which could lead to new or revised information needs on their part. Examples of possible mechanisms are given by the following selection of mechanisms used at Statistics Canada (Fellegi 1996):

- a National Statistics Council to provide advice on policy and priorities for statistical programs;
- professional advisory committees in major subject areas;
- special bilateral liaison arrangements with key federal government ministries;
- participation of the Chief Statistician in policy and program discussions among Deputy Ministers, including access to proposals to Ministers so that the statistical data needs implicit in proposed decisions or new programs can be identified;
- a Federal-Provincial Consultative Council on Statistical Policy, and subsidiary committees on specific subject-

matters, for maintaining awareness of provincial and territorial governments' statistical needs;
- special Federal-Provincial arrangements in the areas of education, health and justice to manage statistical development in these areas of largely provincial jurisdiction;
- meetings with major industry and small business associations;
- feedback through individual users and user enquiries.

These mechanisms are designed to identify gaps in the statistical system – information required by users that is not currently available or good enough for the desired purposes.

3.2 Program Review

The client liaison mechanisms described above will generate user feedback on current programs in addition to information about new and future needs. But periodically some form of explicit program review is required to assess whether existing programs are satisfying user needs, not only in terms of the topics addressed, but also in terms of the accuracy and timeliness of information being produced. Such reviews would utilize information generated by the regular client liaison mechanisms, might also assemble additional data, and would certainly integrate and assess this information to provide a comprehensive picture of how well the program is satisfying client needs.

There are several approaches to such an assessment. An independent expert may be commissioned to consult the user community and make recommendations on program changes. The program area itself may be required to periodically gather and assess the feedback information it is receiving, and prepare a report identifying possible changes to the program. Programs may be required to identify their lowest priority sub-programs so that these can be compared in importance with potential new investments in the same program or elsewhere.

Centrally, the NSO may conduct user satisfaction surveys covering various components of the statistical program, and monitor sales or usage of statistical products. It may also, as a result of its own integrating analytic work, identify gaps or deficiencies in the NSO's products.

All of these approaches have the common feature that, periodically, they call into question, at least on the margins, the continued existence of current programs. They help to identify investment options, both disinvestment from programs no longer relevant, and reinvestment to fill gaps in programs not keeping up with client needs.

3.3 Priority Determination

The final leg of the stool is the process for considering, and acting upon, the information gleaned from user consultations and program review. Since demands will always outstrip the availability of funds, this is a process that requires the exercise of judgement in weighing the diverse needs of different user constituencies. An additional dimension of this process involves recognizing and pursuing opportunities for obtaining new financing to meet high priority information needs, thus reducing the pressure on existing programs to yield resources for reinvestment elsewhere.

At Statistics Canada, the regular annual planning cycle is the core of this process. In this process decisions may be made to invest in feasibility studies in preparation for filling recognized data gaps, to provide seed money to demonstrate how information could be produced with a larger investment, or to invest in improvements to the accuracy, timeliness or efficiency of existing programs. The launching of major new data collection initiatives usually requires resources beyond the means of internal reallocation, so the planning cycle is supplemented by periodic exercises to obtain support and funding from key federal data users for addressing major data gaps (Statistics Canada 1998b). In determining priorities a balance has to be struck between the need for change and improvement, and the need to satisfy the important ongoing requirements served by the core program. In practice, changes from one year to the next are marginal compared to the overall program.

3.4 Monitoring Performance

Measures of performance in the domain of relevance are of two main types. Firstly, evidence that the processes described above are in place is provided by descriptions of the particular mechanisms used supported by examples, if not measures, of their impact. For example, the coverage of consultative mechanisms may be assessed by systematically considering each of the major client or stakeholder groups and identifying the means of obtaining information on their statistical needs. The operation of such mechanisms can be evidenced by reviewing records of their deliberations or consultations. From the program perspective, evidence of periodic evaluation of the current relevance of each program can be provided and the impact of the results of these evaluations can be assessed.

Secondly, direct evidence of relevance may be provided by measures of usage, by client satisfaction results, and by high-profile examples of statistical information influencing or shedding light on important policy issues. Sales of information products and services provide a direct and convincing indicator of relevance. Usage of free products and services, including Internet hits for example, also reflects levels of interest, though the impact of price on usage can be complex and sometimes misleading. Pointing out and publicizing new analytic findings based on NSO data that shed light on important public policy issues can be especially convincing in demonstrating relevance. More generally, regular publication of analytic results in a readable form provides a continuing illustration of the relevance of a NSO's output, especially when republished broadly in the daily press.

Finally, the real changes that the NSO makes in its programs from year to year are a visible reflection of the working of its client liaison and priority-setting processes.

4. ACCURACY

Processes described under relevance determine which programs are going to be carried out, their broad objectives, and the resource parameters within which they must operate. Within those "program parameters" the management of accuracy requires attention during the three key stages of a survey process: design, implementation, and assessment.

4.1 Design

The broad program parameters will not usually specify accuracy targets. They will often indicate the key quantities to be estimated, and the level of detail (*e.g.*, geographical, industrial) at which accurate estimates are needed, but the definition of "accurate" will at best be vague. Nor will they deal at all with tolerable levels of nonsampling error. Indeed, given the multiplicity of estimates and analyses, planned and unplanned, that come from any survey program, it would not be feasible or even useful to try to specify, before design begins, target accuracy levels. The objective of survey design is to find an optimum balance between various dimensions of accuracy and timeliness within constraints imposed by budgets and respondent burden considerations. In this process options that result in different levels of accuracy at different costs, within the broad program parameters, may be considered. The output of the design stage is a survey methodology within which some accuracy targets or assumptions, at least for key estimates and key dimensions of accuracy, will often be embedded. For example, a sample survey may aim to achieve a sampling coefficient of variation for its key estimate below a given threshold at the provincial level, and assume a response rate not less than a defined level. A census design may aim at a specified overall coverage rate, with no key sub-group's coverage falling below some lower specified rate.

The purpose here is not to describe the techniques of survey design that assist in finding optimum designs - that is the subject of the survey methodology literature (amply illustrated by the contents of this journal over its first 25 years!). Here we seek to identify some key management questions that need to be asked to ensure that accuracy considerations have received due attention during the design. We suggest eight primary aspects of design to which attention should be evident.

1. Explicit consideration of overall trade-offs between accuracy, cost, timeliness and respondent burden during the design stage. The extent and sophistication of these considerations will depend on the size of the program, and the scope for options in light of the program parameters. But evidence that proper consideration was given to these trade-offs should be visible.
2. Explicit consideration of alternative sources of data, including the availability of existing data or administrative records, to minimize new data collection. This issue focuses on the minimization of respondent burden and the avoidance of unnecessary collection.
3. Adequate justification for each question asked, and appropriate pre-testing of questions and questionnaires, while also assuring that the set of questions asked is sufficient to achieve the descriptive and analytical aims of the survey.
4. Assessment of the coverage of the target population by the proposed survey frames.
5. Within overall trade-offs, proper consideration of sampling and estimation options and their impact on accuracy, timeliness, cost, response burden and comparisons of data over time.
6. Adequate measures in place for encouraging response, following up nonresponse, and dealing with missing data.
7. Proper consideration of the need for quality assurance processes for all stages of collection and processing.
8. Appropriate internal and external consistency checking of data with corresponding correction or adjustment strategies.

While these eight areas do not cover all aspects of survey design, and consideration of issues does not necessarily result in the "optimum" decision, evidence that these aspects have been seriously considered will be strongly suggestive of sound survey design. In the end, the strength of the survey methodology will depend on the judgements of survey design teams. However, this list of issues provides a framework to guide those judgements and ensure that key factors are considered. Smith (1995) and Linacre and Trewin (1993) illustrate the balancing of these considerations in theory and practice.

Not included in the above list is a ninth area for attention: built in assessments of accuracy. This will be covered in section 4.3 below.

4.2 Implementation

But a good design can be negated in implementation. While a very good design will contain built-in protection against implementation errors (through quality assurance processes, for example), things can always go wrong. From the management perspective, two types of information are needed at the implementation stage . The first is information to monitor and correct, in real time, any problems arising during implementation. This requires a timely information system that provides managers with the information they need to adjust or correct problems while the survey is in progress. The second need is for information to assess, after the event, whether the design was carried out as planned, whether some aspects of the design were problematic in operation, and what lessons were learned from the operational standpoint to aid design in the future. This too requires information to be recorded during implementation (though not necessarily with the same fast feedback as for the first need), but it can also include information gleaned from post-implementation studies and debriefings of staff involved in implementation.

Of course, information pertaining directly to accuracy itself may only be a small subset of the information required by operational managers. But information related to costs and timing of operations is equally important to the consideration of accuracy for future designs.

4.3 Accuracy Assessment

The third key stage of the survey process is the assessment of accuracy – what level of accuracy have we actually achieved given our attention to accuracy during design and implementation? Though we describe it last, it needs to be a consideration at the design stage since the measurement of accuracy often requires information to be recorded as the survey is taking place.

As indicated earlier, accuracy is multidimensional and choices have to be made as to what are the most important indicators for each individual survey. Also each survey produces thousands of different estimates, so either generic methods of indicating the accuracy of large numbers of estimates have to be developed, or the indicators are restricted to certain key estimates.

As with design, the extent and sophistication of accuracy assessment measures will depend on the size of the program, and on the significance of the uses of the estimates. Here we propose four primary areas of accuracy assessment that should be considered in all surveys (Statistics Canada 1992). Other, or more detailed, assessments may be warranted in larger or more important surveys to improve the interpretability of estimates as discussed later.

1. Assessment of the coverage of the survey in comparison to a target population, for the population as a whole and for significant sub-populations. This may mean assessing the coverage of a list frame (*e.g.*, a business register by industry), the coverage of a census that seeks to create a list of a population (*e.g.*, the coverage of a census of population by province or by age and sex), or the coverage of an area sample survey in comparison to independent estimates of the target population (*e.g.*, the difference between sample based population estimates from a household survey and official population estimates).
2. Assessment of sampling error where sampling was used. Standard errors, or coefficients of variation, should be provided for key estimates. Methods of deriving approximate standard errors should be indicated for estimates not provided with explicit standard errors.
3. Nonresponse rates, or percentages of estimates imputed. The objective is to indicate the extent to which estimates are composed of "manufactured" data. For skew populations (such as most business populations), nonresponse or imputation rates weighted by a measure of size are usually more informative than unweighted ones.
4. Any other serious accuracy or consistency problems with the survey results. This heading allows for the possibility that problems were experienced with a particular aspect of a survey causing a need for caution in using results. For example, a widely misunderstood question might lead to misleading estimates for a particular variable. It also allows any serious inconsistencies between the results and other comparable series to be flagged.

The choice of how much effort to invest in measuring accuracy is a management decision that has to be made in the context of the usual trade-offs in survey design. But requiring that, at a minimum, information on these four aspects of accuracy be available for all programs ensures that attention is paid to accuracy assessment across the NSO. It also provides a basis for monitoring some key accuracy indicators corporately. For example, tracking trends in response rates across surveys of a similar type can provide valuable management information on a changing respondent climate, or on difficulties in particular surveys. Regular measures of the coverage of major survey frames such as a business register or an address register also provide information that is important both to individual programs using these frames, and to NSO management. More will be said about the provision of information on accuracy to users under interpretability in section 7.

5. TIMELINESS

Timeliness of information refers to the length of time between the reference point, or the end of the reference period, to which the information relates, and its availability to users. As we have seen, the desired timeliness of information derives from considerations of relevance – for what period does the information remain useful for its main purposes? The answer to this question varies with the rate of change of the phenomena being measured, with the frequency of measurement, and with the immediacy of response that users might make to the latest data. As we have also seen, planned timeliness is a design decision often based on trade-offs with accuracy – are later but more accurate data preferable to earlier less accurate data? – and cost. Improved timeliness is not, therefore, an unconditional objective. But timeliness is an important characteristic that should be monitored over time to warn of deterioration, and across programs to recognize extremes of tardiness. User expectations of timeliness are likely to heighten as they become accustomed to immediacy in all forms of service delivery thanks to the pervasive impact of technology. Unlike accuracy, timeliness can be directly observed by users who, one can be sure, will be monitoring it whether or not the NSO does.

As indicated under accuracy, the explicit consideration of design trade-offs is a crucial component of the management of timeliness in a NSO. Equally, measures described earlier under implementation (see section 4.2) are important in ensuring that planned timeliness objectives are actually achieved. But there are further measures that can be pursued for managing timeliness.

Major information releases should have release dates announced well in advance. This not only helps users plan, but it also provides internal discipline and, importantly, undermines any potential effort by interested parties to influence or delay any particular release for their benefit. Achievement of planned release dates should be monitored as a timeliness performance measure. Changes in planned release dates over longer periods should also be monitored.

For some programs, the release of preliminary data followed by revised and final figures is used as a strategy for making data more timely. In such cases, the tracking of the size and direction of revisions can serve to assess the appropriateness of the chosen timeliness-accuracy trade-off. It also provides a basis for recognizing any persistent or predictable biases in preliminary data that could be removed through estimation.

For ad hoc surveys and new surveys another possible indicator of timeliness is the elapsed time between the commitment to undertake the survey and the release date. This measure reflects the responsiveness of the Agency in planning and setting up a survey as well as its execution after the reference date. But its interpretation must take account of other factors that help to determine how quickly a new survey should be in place – faster is not necessarily better.

For programs that offer customized data retrieval services, the appropriate timeliness measure is the elapsed time between the receipt of a clear request and the delivery of the information to the client. Service standards should be in place for such services, and achievement of them monitored.

6. ACCESSIBILITY

Statistical information that users don't know about, can't locate, or, having located, can't access or afford, is not of great value to them. Accessibility of information refers to the ease with which users can learn of its existence, locate it, and import it into their own working environment. Most aspects of accessibility are determined by corporate-wide dissemination policies and delivery systems. At the program level the main responsibility is to choose appropriate delivery systems and ensure that statistical products are properly included within corporate catalogue systems.

So the management of accessibility needs to address four principal aspects of accessibility. Firstly, there is the need to have in place well-indexed corporate "catalogue" systems that allow users to find out what information is available and assist them in locating it. Secondly, there is the need for corporate "delivery" systems that provide access to information through distribution channels, and in formats, that suit users. Thirdly, the coverage of statistical information from individual programs in corporate catalogue systems and the use of appropriate delivery systems (corporate or in some cases program-specific) by each statistical program has to be managed. Finally, there have to be means of obtaining and acting upon usage and user satisfaction measures for the catalogue and delivery systems.

Given the current rate of technology change, the nature of both catalogue and delivery systems is evolving fast. The traditional printed catalogue that was almost always out of date has given way to on-line catalogues of statistical products, whether printed or electronic, linked to metadata bases in which characteristics of the information can be found. A thesaurus that helps users search for information without necessarily knowing the precise terms used by the NSO is also a crucial component of a catalogue system. Access to the catalogue system can be through the Internet, and users who find what they want can immediately place an order to request the desired information. It is also essential that the NSO's catalogue inter-operate with external bibliographic systems so that users searching outside the NSO are directed to it.

In addition to the structured and exhaustive approach of the catalogue, there are at least two other potential entry points for discovering what data are available. The NSO's official release mechanism in which all newly available data are announced, *The Daily* in the case of Statistics Canada, can provide links to catalogue entries for related products and to sources of more detailed information and metadata. The NSO's main public statistical presentation on its Internet site, known as *Canadian Statistics* in the case of Statistics Canada, can also include similar links to related information and metadata. While these components are not yet fully operational and integrated in many NSOs, this outlines the nature of catalogue systems for the near future.

The Internet is changing the face of delivery systems and promises to become the hub and entry point of such systems for the coming period. But the traditional delivery system of printed publications is still valued by many users, while electronic products on diskette or CD-ROM meet some needs. On-line databases continue to be a central component of a NSO's information delivery systems, whether accessible via the Internet or directly. Among all this hi-tech turmoil, the NSO has to make sure that the public good information needs of the general public continue to be met whether through the media, through public libraries, or through the Internet. The special needs of analysts who require access to microdata present an important set of delivery challenges which are being addressed in several NSOs (see SSHRC and Statistics Canada 1998 for example) but which we will not deal with here.

Increasingly, organizations outside the NSO, both public and private, are playing important roles in improving the accessibility of information produced by the NSO. These organizations may act simply as distributors of data, or may add context or value to NSO data by integrating them with other information or using them in ways that go beyond those that would be appropriate for a NSO. To maximize accessibility, the NSO must be open to opportunities for partnership with such organizations, but must also ensure that its identity as the source of data remains visible and, where appropriate, encourage linkages back to the original, and usually more detailed, data sources held by the NSO.

An important aspect of the accessibility of information is the pricing policy that governs its dissemination. However well-endowed the NSO, resources are limited and the option of providing unrestricted free access to all potential information is not viable. Nor is it desirable because it would destroy a most valuable source of user feedback: measures of real demand for products. A pricing policy needs to balance the desire to make certain basic information freely accessible in the public domain, while recovering the costs of providing specific products, more detailed information, and special requests. Such a policy can promote accessibility, provide a valuable source of information on relevance, and ensure that the resources of the NSO are properly balanced between collecting and processing new data on the one hand, and servicing demands for information from existing data on the other.

Finally, in the process of moving information from statistical programs into the hands of users we have to guard against the introduction of error. At this last hurdle in the process, the wrong information can get loaded into electronic databases; the wrong version of tables can find their way into publications; and enquirers can be given the wrong information over the telephone. Since the potential for these errors occurs at the delivery stage, we include them under accessibility rather than accuracy. Quality assurance systems that minimize the possibility of such errors are a necessary component of these systems.

Since users are the main judge of accessibility, systematic user feedback on catalogue and delivery systems is crucial. This feedback may be derived from (a) automated usage statistics for the various components of these systems, (b) surveys of user satisfaction with particular products, services, or delivery systems, and (c) voluntary user feedback in the form of comments, suggestions, complaints, or plaudits.

Descriptions of cataloguing and delivery systems used by some NSOs can be found in Podehl (1999), Boyko (1999) and by visiting the websites of particular NSOs.

7. INTERPRETABILITY

Statistical information that users cannot understand, or can easily misunderstand, has no value and may have negative value. Providing sufficient information to allow users to properly interpret statistical information is therefore a responsibility of the NSO. Information about information has come to be known as metainformation or metadata. Managing interpretability is primarily concerned with the provision of metadata.

The information needed to understand statistical data falls under three broad headings: (a) the concepts and classifications that underlie the data; (b) the methodology used to collect and compile the data; and (c) measures of accuracy of the data. Essentially these three headings cover respectively: what has been measured; how it was measured; and how well it was measured. Users clearly need to know what has been measured (to assess its relevance to their needs), how it was measured (to allow appropriate analytic methods to be used), and how well it was measured (to have confidence in the results). Since we can rarely provide a profile of all dimensions of accuracy, the description of methodology also serves as a surrogate indicator of accuracy – it allows the user to assess, if they wish, whether the methods used were scientific, objective and carefully implemented. Under each of these headings, more detailed lists of topics can be formulated (Statistics Canada 1992).

There are close relationships between these three headings and other dimensions of quality. The underlying concepts and classifications used are also a prime determinant of coherence (see next section) and the degree to which they conform with national or international standards should be apparent from the metadata. They are also important for the systems that allow users to find out what information is available as described under accessibility (section 6). The description of methodology will reflect the kind of design decisions described under accuracy (section 4.1) and the use of common tools and methods will be relevant to coherence (section 8). The measures of accuracy should reflect the considerations outlined in section 4.3.

That information needed to understand statistical data must be comprehensible is a tautology worth stating. The NSO has to make a particular effort to ensure that the information provided under these headings is written in the users' language and not in its own internal jargon. Otherwise it fails on interpretability twice over.

To manage the interpretability dimension of quality, we suggest three elements need to be in place. The first is a policy on informing users of the basic information they need to interpret data. This policy would prescribe what information should be provided with every release of data, and in what form it might be provided. The second element is an integrated base of metadata that contains the information needed to describe each of the NSO's data holdings. Typically, this metadata base would contain more than the minimum required by the policy. Thirdly, there is a need for direct interpretation of the data by the NSO. With each major release, there should be some commentary that focuses on the primary messages that the new information contains. Directed particularly at the media, such commentary increases the odds that at least the first level of interpretation to the public will be correct. Conversely, the NSO should answer or refute serious misinterpretation of its data.

Interpretability is perhaps the one dimension of quality where the NSO should aim to do more than the user is asking. There is an element of user education in the provision of metadata. Spreading the message that all data should be used carefully, and providing the information needed to use data with care, is a responsibility of the NSO that goes beyond simply providing what users seek.

The assessment of success in the area of interpretability requires measuring compliance with the policy proposed above, and seeking user feedback on the usefulness and adequacy of the metadata and analysis provided.

8. COHERENCE

Coherence of statistical data includes coherence between different data items pertaining to the same point in time, coherence between the same data items for different points in time, and international coherence. The tools for managing coherence within a NSO fall under three broad headings.

The first element is the development and use of standard frameworks, concepts, variables and classifications for all the subject-matter topics that the NSO measures. This aims to ensure that the target of measurement is consistent across programs, that consistent terminology is used across programs (so that, for example, "educational level" means the same thing whether measured in a Census of population or from school records), and that the quantities being estimated bear known relationships to each other. The realization of this element is normally through the adoption and use of frameworks such as the System of National Accounts and standard classification systems for all major variables. The issue of international comparability is addressed by considering the adherence of the standards adopted to international standards where these exist. Policies are required to define program responsibilities for ensuring that data are produced according to the standards adopted.

The second element aims to ensure that the process of measurement does not introduce inconsistency between data sources even when the quantities being measured are defined in a consistent way. The development and use of common frames, methodologies and systems for data collection and processing contribute to this aim. For example, the use of a common business register across all business surveys ensures that differences in frame coverage do not introduce inconsistencies in data (there are other reasons for using a common business register too); the use of commonly formulated questions when the same variables are being collected in different surveys serves to minimize differences due to response error; the use of common methodology and systems for the various processing steps of a survey, especially edit and imputation, helps to ensure that these operations do not introduce spurious differences in data. All of these arguments apply across occasions of a particular survey, as well as across surveys.

With the first two elements we attempt to ensure that we do not build into the design or implementation of statistical programs any unjustified inconsistency. The third element deals with the results of this attempt and focuses on the comparison and integration of data from different sources. Some integration activities are regular and routine, *e.g.*, the integration of data in the national accounts, benchmarking or calibration of estimates to more reliable control totals, seasonal adjustment of data to facilitate temporal comparisons. Other activities are more exploratory and ad hoc. The confrontation of data from different sources, and their subsequent reconciliation or explanation of differences, is an activity that is often needed as part of pre-release review or certification of data to be published. Feedback from external users and analysts of data that point out coherence problems with current data is also an important component of coherence analysis. Some incoherence issues only become apparent with the passage of time and may lead to historical revisions of data.

To assess success in achieving coherence one can identify three broad sets of measures corresponding to the three elements described above. The existence and degree of use of standard frameworks, variables and classification systems; the existence and degree of use of common tools and methodologies for survey design and implementation; and the incidence and size of inconsistencies in published data. Within this latter category, one might include, for example, monitoring the residual error of the national accounts, the closure error in population estimation, or the size of benchmarking adjustments in major surveys.

9. OVERALL MECHANISMS

In reviewing each dimension of quality we have identified mechanisms which we believe to be important for the management of quality within a NSO. Some of these mechanisms lead to measures that have to be taken or followed by each individual statistical program within the NSO. Others lead to corporate-wide systems which all programs use, or to which they contribute information. In this section we extract what we consider to be the five major components or subsystems of a quality management system within a NSO.

The user liaison subsystem consists of the series of mechanisms that serve to keep the NSO in touch with its primary user groups. It provides information about current and anticipated information needs, adequacy of current products, and advice on priorities. It plays a key role in assuring the relevance of the NSO's output.

The corporate planning subsystem takes the information coming in from the user liaison system, together with assessments and internal knowledge of program strengths and weaknesses, to identify where program reductions or investments should be made. It sets the program parameters for all programs, and therefore has a direct impact on the relevance, accuracy and timeliness achievable by statistical programs. Through its funding decisions on infrastructure programs, it also influences directly the accessibility, interpretability and coherence of statistical outputs. It must be overseen by the NSO's senior management committee. Funding decisions depend on a robust cost reporting system that accurately captures the component costs of statistical programs.

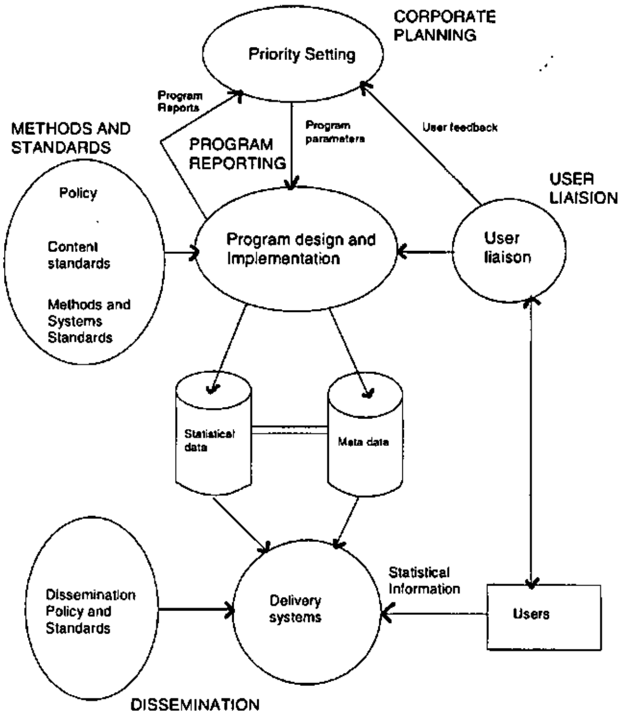
The methods and standards subsystem establishes the policies and guidelines that govern the design and implementation of statistical programs, including both content and documentation standards, and standards for the methodology and systems used. It is key to achieving coherence and interpretability across statistical outputs, and to the optimization of accuracy and timeliness within programs. Its management must involve senior representation from across the NSO through a management committee.

The dissemination subsystem establishes the policies and guidelines, and puts in place the corporate systems, for delivering information to users. This includes the management and delivery of the metadata needed by users to search and access the NSO's data holdings. It is the key determinant of the accessibility and interpretability of the NSO's data. Its management too must involve senior representation from across the NSO.

Last but not least is the program reporting subsystem. Whatever the level of corporate emphasis on quality, it is within the individual statistical programs that quality is built in to the products. Within the constraints and guidance provided by corporate policies and guidelines, individual programs have to make informed trade-offs and decisions that will influence quality in all its dimensions. Within programs, evaluation and analysis of data provides a first assessment of the accuracy and coherence achieved. It is individual programs that have to defend their accuracy and timeliness records to users. A system for regular reporting by programs to management on their achievements in the different domains of quality provides an essential management input, not only for current monitoring, but more importantly as an input to the corporate planning subsystem where decisions on future investments are made.

Diagram 1 provides a simplified sketch of the relationships between these five subsystems, or key functions, necessary to the management of quality in a NSO. The subsystems are not organizational units. Indeed, the nature of most of them is that they must involve a cross-section of staff from across the NSO in order to build a corporate consensus on the appropriate policies and standards to be followed.

MANAGING QUALITY IN A NSO



Speaking of staff, the diagram omits the crucial role of staff in all of these subsystems. A NSO is heavily dependent on a strong cadre of "knowledge-workers" covering a wide range of disciplines. As we have seen, professional expertise and judgement are required in many aspects of the design, analysis and evaluation of statistical programs. Competent staff are required to execute all phases of statistical programs with attention to the assurance of quality. Surrounding the subsystems described, we should envisage a human resources subsystem that aims to ensure that the NSO has at all times a well-trained, motivated and versatile staff capable of meeting the challenges facing the NSO. In particular, they need to have an appreciation of the importance of satisfying client needs through the management of all dimensions of quality. For one approach to a human resources subsystem see Statistics Canada (1997).

10. CONCLUSION

One object of this paper has been to put the statistician's traditional concern for accuracy into a broader context. Accuracy is important, but without attention to other dimensions of quality, accuracy alone will not satisfy users. Nor for many users is it the most important consideration. Trying to look at quality from a user perspective may help in solving the inevitable trade-offs between accuracy and other dimensions of quality.

This broader view of quality also helps to link together several key activities within a NSO as contributors to the management of quality. Training activities within the NSO can take advantage of this linkage to develop a broader understanding among employees of how different activities within the NSO fit together or complement each other - and particularly of why their own work is so important.

This broader view also helps to reinforce the importance of analysis within a NSO. Analysis has been mentioned as a means of demonstrating relevance, as a means of checking accuracy, as a means of improving interpretability, and as a means of testing coherence. And that list excludes the basic role of analysis in adding to the information content of statistical outputs.

For the future, more can be done to refine the concept of quality in a NSO and to improve quality management. Within the narrower domain of accuracy, there is still more room for work on the control and measurement of non-sampling errors. With the increasing reliance on administrative data, more systematic study of the attributes of data from these sources will be required. The growing interest in longitudinal surveys, sometimes linked with administrative data, raises the need to manage accuracy issues arising in such surveys. Finally, the growing emphasis on the combination of data using integrating frameworks calls for more attention to the quality attributes of data resulting from such manipulations.

ACKNOWLEDGEMENTS

The author recognizes valuable comments from many colleagues in Statistics Canada and very helpful suggestions from the Editor and a referee.

REFERENCES

BOYKO, E. (1999). Statistical meta-data in context: an overview of statistical meta-data and related meta-data systems. Paper prepared for the 1999 Conference of European Statisticians, UN/ECE Work Session on Statistical Metadata, Working Paper No. 17, Geneva, September 1999.

COLLEDGE, M., and MARCH, M. (1997). Quality policies, standards, guidelines, and recommended practices at national statistical agencies. *Survey Measurement and Process Quality*, 501-522, New York: John Wiley.

COLLINS, M., and SYKES, W. (1999). Extending the definition of survey quality. *Journal of Official Statistics*, 15, 1, 57-66.

de LEEUW, E., and COLLINS, M. (1997). Data collection methods and survey quality: an overview. *Survey Measurement and Process Quality*, 199-220, New York: John Wiley.

DIPPO, C. S. (1997). Survey measurement and process improvement: concepts and integration. *Survey Measurement and Process Quality*, 457-474, New York: John Wiley.

EUROSTAT (1998). Papers prepared for the 1998 DGINS meeting are available from EUROSTAT, Luxembourg. See also SIGMA, The Bulletin of European Statistics, 03/1998, Quality in Statistics, published by EUROSTAT.

FELLEGI, I.P. (1996). Characteristics of an effective statistical system. *International Statistical Review*, 64, 2.

HANSEN, M.H., HURWITZ, W.N., and PRITZKER, L. (1967). Standardization of procedures for the evaluation of data: measurement errors and statistical standards in the Bureau of the Census. *Bulletin of the International Statistical Institute, Proceedings of the 36th Session*, 49-66.

LINACRE, S., and TREWIN, D. (1993). Total survey design – application to a collection of the construction industry. *Journal of Official Statistics*, 9, 3, 611-621.

LYBERG, L., BIEMER, P., COLLINS, M., de LEEUW, E., DIPPO, C., SCHWARZ, N., and TREWIN, D. (Eds.) (1997). *Survey Measurement and Process Quality*, New York: John Wiley.

PODEHL, W.M. (1999). Data base publishing on the internet. *Statistical Journal of the United Nations Economic Commission for Europe*, 16, 145-153.

SMITH, T.M.F. (1995). Problems of Resource Allocation. *Proceedings: Symposium 95, From Data to Information – Methods and Systems*, Statistics Canada, 107-114.

SSHRC and STATISTICS CANADA (1998). Final Report of the Joint Working Group of the Social Sciences and Humanities Research Council and Statistics Canada on the Advancement of Research using Social Statistics, December 1998.

STATISTICS CANADA (1992). Policy on Informing Users of Data Quality and Methodology, April 1992. Policy Manual 2.3

STATISTICS CANADA (1997). Human Resources Development at Statistics Canada, November 1997. Internal document.

STATISTICS CANADA (1998a). *Quality Guidelines*. Third Edition, Catalogue No. 12-539-X1E, Statistics Canada.

STATISTICS CANADA (1998b). Statistics Canada's Corporate Planning and Program Monitoring System, October 1998. Internal document.

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (1999). Papers prepared for the 1999 Plenary discussion. Statistical Division, UNECE, Geneva (www.unece.org/stats/documents/1999.06.ces.htm).

UNITED NATIONS ECONOMIC AND SOCIAL COUNCIL (1994). Report of the Special Session of the Statistical Commission, New York, 11-15 April 1994, E/1994/20.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 151-160
Statistics Canada

Administrative Records and Census Taking

FRITZ SCHEUREN[1]

ABSTRACT

The shift in the use of administrative records from an incidental role in census applications to an essential one is now well along in many European countries. The challenges are greater in Canada and the U.S., as this paper discusses. Progress in the U.S. in developing a modified administrative census paradigm is dealt with in some detail and contrasts made to what has already been done elsewhere, notably in Canada. A research agenda is set out and some connections made across a whole gambit of U.S. census-connected statistical programs – including current surveys, intercensal estimates, and the measurement of the census undercount. Privacy concerns are prominent among the issues that are addressed. The role of low cost computing and advanced record linkage software also are given their due. The changing status of central statistical agencies as the information age advances is also touched on.

KEY WORDS: Record linkage; Privacy; Surveys and intercensal estimates.

1. INTRODUCTION

The use of tax and other governmental administrative records in census-taking turns out to be quite old, even though most of the real advances have occurred in the last 25 years or so – that is, roughly during the same time span as the publication of *Survey Methodology*.

The introduction of modern sampling – or the representative method, as Kiaer called it (Bellhouse 1988) – was tied, it seems, to the matching of samples of Norwegian tax records to the census of 1890 in Norway (Johnson and Kotz 1997). Of course, the mathematics of Kiaer's approach, rather than the particular application, was the focus of much of the later work (and remains the usual focus in *Survey Methodology*). This was appropriate, given that administrative records were often inaccessible and hard to use.

Over time, though, there has been legislation (like the Canadian Statistics Act) that has made access to administrative records routinely possible by Central Statistical Offices in many countries. Advances in computing and recordkeeping in government and elsewhere, while bringing new problems, have certainly also made administrative records easier to use for statistical purposes and this trend seems likely to continue or even accelerate (*e.g.*, Kenessey 1994).

Traditional censuses have been replaced in whole or in part in some of the Nordic countries by administrative records (*e.g.*, Myrskla 1991; Thomsen and Holmoy 1998). This has not occurred in Canada or United States – partly because of the nature of the administrative records available and partly because of the sheer size and complexity of the undertaking. In fact, even in one of the most ambitious North American administrative record census (ARC) proposals (*e.g.*, Alvey and Scheuren 1982), the complete elimination of conventional census-taking was not advocated. Rather a mixed mode approach was suggested.

The current paper attempts to recount briefly the "state of the play" on administrative record census proposals (See Steffey and Bradburn 1994 for an additional perspective). The paper focuses mainly on the U.S. but with Canadian parallels (Leyes and Elsl-Culkin 1994). In some respects this is a follow-on or update to a piece in *Survey Methodology* ten years ago (Scheuren 1990).

Before going into details, it may be worth providing a context to the changes needed to achieve some form of an administrative record census. First, it may make sense to discuss the nature of scientific revolutions generally. Bellhouse (1988) cites Kuhn (1970) in this regard relative to sampling itself. Scheuren (1990) also drew independently on Kuhn concerning ARC ideas.

For example, it was not really until the paper by Neyman (1934) – aided by Sukhatme (1935) and, through Deming's advocacy, plus the all important paper by Hansen and Hurwitz (1943) – that Kiaer's randomization-based approach to the representative method might be argued to have been accepted.

At least in North America, ARC ideas may not yet have found their Neyman. Still, there have been many hard-won successes to celebrate and with the fuller emergence of enabling technologies (like record linkage and low-cost computing), the shape of the future can be characterized as encouraging.

Organizationally, the present paper is divided up into 8 sections, beginning with this introduction (section 1). Section 2 sets out some ARC background and section 3 develops a few assumptions about issues that go beyond the operational feasibility of an ARC. The rest of the paper consists of suggestions in four areas: the 2000 U. S. Census (section 4), the intercensal population estimates program for the coming decade (section 5), the current surveys program (section 6), and the planning for the 2010 U.S. Census (section 7). There is also a concluding section (section 8)

[1] Fritz Scheuren, The Urban Institute. Mail: 1402 Ruffner Road, Alexandria, VA, 22302, U.S.A., e-mail: scheuren@aol.com.

that discusses priorities. Finally, some references are provided.

2. BACKGROUND ON ORIGINAL ARC PROPOSAL

In Europe, several countries are quite far along in developing administrative record censuses, having begun back in the 1970's (*e.g.*, Jensen 1983; Jabine and Scheuren 1987; Redfern 1989; Blum 1999). Those countries are much smaller and differ in many other ways from the U.S. or Canada – especially in the social contract that underlies census-taking. What they have done, therefore, is hard to apply directly. Still, their pathbreaking efforts have much to teach.

The original idea for a partial ARC in the United States was first made publicly at an American Statistical Association meeting in 1982 (Alvey and Scheuren 1982). The work of John Leyes and Doug Norris at Statistics Canada was one of the inspirations for that proposal. Basically, the paper advocated research on how –

> To link U.S. Internal Revenue Service (IRS) tax return data to wage and retirement earnings, unemployment compensation records, and U.S. Health and Human Services (HHS) administrative files to obtain a "bare bones" population census.

The key element here was researching a partial replacement for a conventional census – not to completely replace it. The administrative records were, moreover, limited to those already legislatively available in whole or in part to the U.S. Census Bureau. Speculations were offered that some administrative system changes might be possible to accommodate an ARC use; even so, this proposal never contemplated "content-wise" that the resulting ARC would be much more than a bare bones population count.

The anticipated coverage of an ARC was believed, however, to be good but not treated as perfect. In fact, the ARC proposal always assumed some form of sampling to adjust the population for completeness. The prediction was made that the proposed ARC would cover well over 95% of the population covered in a conventional census. The 1993 and 1998 papers by Sailer and his colleagues confirmed this conjecture (Sailer, Weber, and Yau 1993; Czajka, Moreno, and Schirm 1997; Sailer and Weber 1998).

The bare bones aspect might be best illustrated by the fact that no provision was made for the housing census that is conducted along with the current U.S. population census. Housing would have to be dealt with in some other way. Among the weaknesses of the proposal, acknowledged at the time, was the quality of the race data in administrative records and the problem of having mailing rather than actual residential addresses.

Considering these limitations, why proceed? Well, the ARC originally proposed not only would reduce the cost and burden of a decennial census, but has the potential for producing a total population count more frequently than every 10 years. It also might provide improved coverage for some of the populations traditionally undercounted in a decennial census. Moreover, Bye's 1997 work (Bye 1997), plus his recent detailed look at Social Security Administration data on race and ethnicity (Bye 1998a and 1998b), put these weaknesses into perspective and go a long way to suggesting how they could be overcome or at least lived with (See also Bye 1999.)

The most important point about the proposal was that it advocated research towards a potential ARC 10 or even 20 years down the road. Implementing an ARC was not proposed, although some of the reaction raised this concern. Privacy and confidentiality aspects were prominently mentioned in the proposal as also requiring research.

There is no need here to carry the story forward in detail from the original Alvey and Scheuren paper until now. That has been done elsewhere (Scheuren 1995a). What is important to mention is the shift in the tone of the research over the years, from "proving" an ARC could not work to trying to find ways that it might. Bye, for example, in his excellent report to the Census Bureau (Bye 1997), fully spelled out a way to implement such a census. While it has many researchable elements, Bye's approach demonstrates that the idea is operationally feasible.

3. ASSUMPTIONS

Certainly the technology of record linkage and the widespread availability of massive fully-computerized record systems make the creation of alternatives to conventional censuses possible outside the U.S. Federal sector. State governments have incentives to be sure that every resident is counted (Biskupick 1998) and certainly could construct partial ARCs using their own record systems. The motivation to challenge the Census Bureau monopoly is definitely present with the devolution of Federal activities to the states and the financial incentives involved in Federal grant programs. Nearly $200 billion in Federal aid is distributed annually based on population.

3.1 Massive Data Sets

The mass marketers and telephone survey organizations also have extensive data systems that might be tapped into. Private data sources unheard of a few years ago (*e.g.*, even from grocery chains!) are expanding rapidly and extensive statistical use of these private sources is already occurring (National Academy of Sciences 1996). With the worldwide revolution in electronic recordkeeping practices, there will be many new entrants in the emerging information industries. The "hurdle" price has been lowered and the value of information has been growing.

Some recent work done for the State of Connecticut might be worth illustrating the general points just made. In White, Mulrow and Scheuren (1999), the authors describe an effort commissioned by the State of Connecticut to use state administrative records to improve Connecticut's jury selection system. It is important to note at the outset that the goal of that work was not to do an ARC. Still the exercise has a lesson in it about the ease with which a partial ARC could be developed for a state.

Formerly, Connecticut employed voter registration and motor vehicle files with a labor-intensive process to unduplicate the two systems, so as to form a list from which to draw potential jurors for duty. The new effort, described in White *et al.* (1999), involved employing probability-based linkage technologies (Jaro 1989) with four state-level files: the two mentioned already, plus the State's income tax file and the State's unemployment file. The files were all created in early 1998.

To evaluate the Connecticut linked data, a comparison was made to 1996 Census Bureau population projections by township, brought forward to 1997. The administrative record population coverage obtained by the combined file was surprisingly good, given that an ARC was not the goal. In fact, the linked administrative record counts by township were highly correlated with population projections. The simple correlation was $\rho = 0.946$. When four of the 169 townships are removed as outliers, the correlation went up to $\rho = 0.977$.

3.2 Privacy Considerations

Of course, privacy assumptions bear on direct use of administrative lists and on linkages across them. Obviously, ARC considerations about personal privacy will impact linkages of data from different sources for census purposes. In 1985, early results of the privacy research on linkage issues were presented (Scheuren 1985), followed by a great deal of other work, notably by the Census Bureau – reported on, for example, by Gates and Bolton (1998), Gates (1999) and Singer (1999).

It looks reasonable, despite concerns, such as those in Scheuren (1997), that a careful introduction of greater and greater linkage will succeed in gaining wide acceptance as a policy. In fact, Statistics Canada is already experimenting with this now through their Survey of Financial Security, where respondents are given an opportunity to authorize access to tax and pension records instead of responding to selected survey questions (Statistics Canada 1999). In that survey, they are finding very high acceptance of the idea. This reference is just an example of the success that has already been achieved in direct uses of administrative records in Canadian surveys. For example, the option of accessing tax records has been standard in the Canadian Survey of Labour and Income (SLID) since May 1995. (See Statistics Canada 1993-1996.)

While, in the United States, perhaps a sixth of the population will object, their views may not be listened to. Despite this, it appears likely that there will be no outcry and the "taking" of these privacy rights will proceed with little incident. Fellegi (1997), in his opening address at the 1997 International Record Linkage Conference, gave a sound analysis of this possibility.

3.3 Access Considerations

For the Census Bureau to do an ARC would require new legislation to mandate cooperation by various government agencies with the Bureau. Currently, except for the IRS, the Bureau may receive administrative data if other agencies choose to provide them; but, unlike the Statistics Act in Canada, there are no laws that require agencies to cooperate. Continuing this arrangement, of course, would be untenable if the Census Bureau were to try an ARC.

The development and enactment of such legislation would provide the opportunity for a public debate on ARC ideas, something that must occur before an ARC could be done. In any case, legislation is required that would mandate cooperation with ARC research; otherwise, the Bureau may never get to do the required preparatory work. This suggests legislation "now," if the Census Bureau is to prepare for an ARC in 2010.

3.4 Technological Advances

The assumption is that there will also be continuing advances in record linkage techniques, led by Bill Winkler at the U.S. Census Bureau and Martha Fair, among others, at Statistics Canada. The data mining "craze" can be anticipated to lead to a very wide dissemination of these techniques. Large privately-held data sets will be increasingly combined and in an increasingly statistically satisfactory way. Tied to this growth will be a realization, as in Scheuren and Winkler (*e.g.*, 1993 and 1997), that the goal of linkage is not mainly the matched data, but a way to combine disparate sources to produce information otherwise unattainable because of cost.

There will continue to be an expansion of access to and uses of improving Geographic Information System (GIS) software, especially in small area estimation applications, both within and outside of Central Statistical Offices. We are entering a new "data-dense" world, where the amount of information available geographically is exploding. Much of this will be estimated, but the overall quality will be superb. Increasingly, isolated estimates (as in Schaible 1996) will be replaced by sets of interlocked covariates that are coherent together. In all likelihood, market forces will drive this. The impact of cheaper and more powerful computing will mean that the handling of very large files and burdensome computations will not be seen as barriers, even in government – albeit there will be a lag in the public sector.

If these scenarios happen, the world of high cost data gathering (like a conventional census) will increasingly be replaced by a world of frugal reuse of data – often automatically obtained (*e.g.*, as predicted in International

Statistical Institut (1994). Widespread reuse applications will spur even better techniques and, combined with competition and cheap computing, will reduce greatly the power of data producers, including Central Statistical Offices.

4. CENSUS 2000 SUGGESTIONS

To develop an administrative record census, much research is clearly needed. This section sets out suggestions for administrative record research to be done as part of the 2000 Census in the United States. These are grouped into process (section 4.1) and content (section 4.2) suggestions.

4.1 Process Observations

The 2000 Census "kicks off" a decade of potential activity in getting ready for Census 2010. The observations made here on these possible activities fall under four headings: acquiring more administrative data, strengthening the safeguards on use, building cooperative arrangements for staff exchanges, and establishing a precedent of modifying existing administrative systems to enhance their information uses.

4.1.1 Data Acquisition

There certainly is a history of greater cooperation at census time by other government agencies. While there were many complications, it is no coincidence that the first IRS Individual Master File extract that the U. S. Census Bureau received was obtained for income year 1969 of returns filed in the decennial census year 1970. The occasion of the 2000 Census should be used (and has), therefore, to advance the 2010 agenda, by acquiring data and exploring how to use them to develop an ARC.

The Census Bureau's recent precedent in obtaining the full Social Security Number (SSN) application or Numident file from the Social Security Administration (SSA) is a particularly important example of the kind of acquisition needed, since the file contains age and other demographic data items on all persons who have SSNs. As Prevost and Leggieri (1999) discuss, there are many efforts underway which have led to the Census Bureau obtaining still more Federal record systems.

Obtaining pilot access to state program records for the medically indigent (Medicaid) should be a priority; this is so despite the quality issues that such systems have. Make no mistake, however; a wholesale acquisition policy could be perceptually dangerous (*i.e.*, violating the privacy assumption mentioned in section 3). Only systems for which there are clear, sustainable research objectives (and financial support) should be sought.

It is important to point out that a census requires not only a full count of the population but must include correct geographic location at a point in time. For apportionment, state-level geography is required; for redistricting, geographic location well below the state-level is required. The implication of this for data acquisition is twofold. First, the administrative record files must attempt collectively to cover the "entire" population. Second, the files must provide good information on low-level geographic location at chosen points in time. Sometimes, even, files should be obtained just because they provide better geographic location for some part of the population (see Bye 1997 for more details.).

IRS acquisitions might be of two types: small incremental additions, as well as acquisitions of full-scale tax files already being received by the Census Bureau. The small additions are technical and procedural, involving working level staffs; the larger acquisitions have policy elements and need a different approach – with involvement at the highest level.

Regularly since 1969, the Census Bureau has obtained an extract from the IRS Individual Master File system. Late returns, not filed in time for that extract, are becoming increasingly important and might be added to the data from IRS. Second, the prior year returns should also be obtained and introduced into the longitudinal samples recommended in sections 5 and 6 below. Marginally increasing item content to include more types of income is also suggested. Obtaining all or a large sample of information master file documents electronically is recommended. Getting all wage and social security information records, as has been done, is an exceedingly good start and certainly seems a plausible compromise for 2000, but interest and dividend records are important too.

In any case, a major effort should be made to provide budget support in non-census years for sustaining this system – a problem that the administrative records program at the Census Bureau has had historically. It can be argued, until recently in fact, that the Bureau already has had more administrative record data than it had resources and people to use fully.

4.1.2 Physical and Perceptual Security

Clearly enhanced physical security of administrative data goes hand in hand with more data acquisitions. The Census Bureau recently established a secure restricted access environment for its demographic administrative records (Clark and Gates 1999). However, the Census Bureau must not stop there. An outside auditing firm should be hired to test the new physical security. In fact, such efforts should be an ongoing part of the Census Bureau's new data steward role for administrative records. Assuring protection of the data is critical to the success of an ARC. It is important to recognize that linked administrative record databases are inherently more valuable than individual agency files; employees are subject to more temptation or at least the suspicion of being vulnerable. In fact, violations by IRS employees which came to light several years ago (see Scheuren 1995b), led to anti-browsing legislation specific to tax data. The Census Bureau must take every precaution to enforce such rules for all of its administrative records. There is also a need to keep up public opinion survey research and conduct more focus groups with the various stakeholders and the general public, as well as with the Bureau's own employees. The cost of maintaining massive administrative record systems involves both physical and perceptual maintenance of data security. And neither of these comes with a small price tag.

4.1.3 Building Cooperative Arrangements for Staff Exchanges

Human capital improvements are also key to any administrative record initiative. Professional statisticians outside the administrative agency too often think of just the data products they obtain rather than the system as a whole. Some would argue that the unfortunate phrase, "exploiting administrative data," grows out of this narrow (and denigrating) view. Whether the phrase is unfortunate or not, it reflects the hunter-gather phase in the use of administrative records for statistical purposes. That age is ending.

The real (or new) goal should be to turn "administrative systems into information systems" (Scheuren and Petska 1993). This means we need to move, continuing the analogy, to the next or agricultural stage in the use of administrative records.

One way this new phase might be speeded up would be through something like an American Statistical Association fellows program. A sabbatical might be paid for by the Census Bureau and offered to operating administrative agency staff – perhaps from around the world. This could involve having IRS, SSA, and other administrative record stewards in residence at the Census Bureau for short periods. Among the goals would be to give them an understanding of the importance of the information services that their administrative systems made possible. A by-product would be the invaluable insights the administrative agency staff could provide regarding assumptions about and use of their data for statistical purposes.

More important still could be the reverse exchange – Census Bureau staffers going to work at the operating agency for an extended period of time. Unlike in Canada, which has a great deal of professional migration into and out of Statistics Canada to administrative agencies, the U.S. has very little. Anyway, more is needed. Think of the stimulus that this could give the statistical imaginations of the individuals sent. Deming talked about the need for systems thinking (Deming 1986). How better for people to obtain such thinking in connection with administrative systems than by such an experience, repeated periodically every few years.

4.1.4 Establishing a Precedent for Modifying Existing Administrative Systems to Enhance Information Uses

Improving the statistical data products derived from administrative systems can be achieved in many ways. One is to add an item (and the associated burden) to an existing administrative system. Naturally, this is a two-edged sword. Obtaining residential addresses on tax returns, for example, as was done in 1981, would be an obvious example; however, see Bye (1997), where another – and perhaps better – approach is advocated that would involve a direct followup for addresses that are clearly not residential.

Another potential addition to the tax return might be a conventional (or landline) residential telephone number. While the growing use of cellular phones may make such numbers of only temporary value, they still might be worth obtaining. In the U.S. at least the shift to cellular has not been accompanied by the abandonment, yet, of earlier technologies. In any case, it can be predicted that the administrative uses of these numbers could more than pay for their value as a statistical tool in record linkage during the census and later on in an ongoing survey program. Moreover, for listed numbers, there would be a valuable check on the address.

While probably very hard to accomplish, changing third party wage reports (IRS Forms W-2s; T-4s are the Canadian counterpart) so that they have the date of the last pay period on them, would be an enormously valuable addition from an ARC perspective. The addition of the date of the last pay period covered could remove much of the ambiguity associated with multiple addresses on such documents. Of course, accessing the quarterly unemployment system wage records, through the U.S. Bureau of Labor Statistics, might be even better and would not increase existing burden.

These suggestions, while perhaps feasible, will require a great deal of work to implement, since there are many other stakeholders and costs to consider. One observation, which Bye included in his 1997 report on a possible ARC, is to obtain from the U.S. Social Security Administration the mailing address files that are used by them to send SSNs back to the parents of newborns. Here the burden is slight and the value sizable, since it would give access to current addresses for families with new borne children.

Clearly, some tradeoffs are easier to make than others. It is essential, whenever possible, is to find ways that better join an information purpose to an existing administrative one – thereby obtaining something of value for everyone.

4.2 Research Suggestions

There are many worthy research ideas that could be recommended. Two important ones are (1) obtaining SSNs on the post-censal quality check samples to be drawn, so that a triple-systems estimate can be obtained of the undercount; and (2) producing a limited ARC estimate during 2000 for cross-checking with the official "counts." For these to be fully effective the results from both are needed on the same schedule as the official Census Bureau counts and undercount adjusted estimates – due in December 2000 and March 2001 respectively.

4.2.1 Triple Systems Estimation

It has long been advocated (*e.g.*, Scheuren 1995a) that a triple-systems estimate be attempted (Zaslavsky and Wolfgang 1993). The three systems would be the quality check sample, the census itself, and an amalgam of unduplicated administrative records. For triple systems estimation to succeed, all the matching needs to be of high quality. Without SSNs obtained in the quality check post-enumeration survey, the matching to administrative records will be a lot harder and, for doubtful cases, perhaps fatally ambiguous. People with multiple addresses and common names would be particularly challenging in the absence of SSNs.

4.2.2 Concurrent Partial State Level ARC

Even without attempting a triple systems approach, a concurrent limited ARC has potential in any post-census review. The need to have an immediate check on the statewide counts could be accomplished using the methods employed twice now by Sailer, Webber and Yau 1993 and Sailer and Weber 1998 and could be done quickly, if given a high enough priority. The needed IRS administrative records are expected to be essentially in place by the early fall of 2000 and could be processed by the Census Bureau on a flow basis.

Specifically, it is recommended that the Census Bureau receive its normal IRS Individual Master File extracts monthly, so matching can begin early. Information documents on wages earners and social security recipients could also be received on a flow basis from the Social Security Administration (even before being compiled at IRS). Many of the decennial census misses that are in the IRS data bases could well be earned income tax credit (EITC) recipients who may move. Continuous matching and sample checking will be key for finding such individuals. Certainly those EITC filers who use refund anticipation loans will need extra attention, if followup is going to be successful.

There are, of course, many other worthy 2000 Census research ideas that might lay the groundwork for an eventual U.S. Administrative Record Census (see Prevost and Leggieri 1999). The two mentioned above seem, however, far and away the most important. For a recent paper on the use of administrative records in the Canadian Census, see Carter and McClean 1996.

5. INTERCENSAL IMPLICATIONS

Paradoxical as it may sound, to make revolutionary advances in the use of administrative records an evolutionary approach is needed – especially in the intercensal estimates program.

5.1 Annual Administrative Record Portion of ARC

First of all, the Census Bureau should continue annually, on at least a sample basis, the ARC estimation of state totals mentioned in section 4.2. Eventually, depending on data acquisitions and funding, these could be enlarged and deeper geography obtained.

5.2 Large Longitudinal Administrative Sample

Second, large longitudinal administrative record samples should be mounted. Following the Canadian example, the Census Bureau could begin with a straightforward longitudinal sample of tax return records, matched to the U.S. Social Security Administration's Numident file, containing demographic information for all those with SSNs. Statistics Canada has long had a 10% longitudinal sample of T1 returns (Leyes and Elsl-Culkin 1994), which in the U.S. would translate into a 1% sample, given relative country sizes. In fact, tying this longitudinal sample to the U.S. Social Security Administration's 1% Continuous Work History Sample (CWHS) could give it a very long (time) footprint, indeed.

Eventually, this longitudinal sample might be extended across other Federal administrative systems (at IRS and SSA, but perhaps elsewhere too). A caution, though. The chore of matching changing administrative units over time may require more resources and patience than might be anticipated and so should proceed incrementally with smaller efforts, say the 0.1% CWHS for example – as has already been partially implemented (Czajka and Walker 1989).

5.3 Integrated Administrative Statistical Sample

Scheuren (1979) has a much more ambitious 20-year old proposal for a set of interlocking administrative samples. Perhaps this should be re-examined and updated. His ideas involved both standalone efforts and efforts potentially supportive directly of traditional intercensal and current survey programs. Unlike the basic ARC concept, they would have expanded item content as their main objective, rather than complete or near-complete population coverage. They could also be used as starting points for various current survey efforts, as is the case now in the dual frame Survey of Consumer Finances (SCF) mounted by the Federal Reserve Board (Kennickell and Woodburn 1997).

5.4 Transaction-Based System

Fourth, for the long term, the current intercensal administrative records program should move, to the extent it can, from annual data systems with year by year matches towards direct transaction-based adjustments of the administrative counts. Consider, for example, an effort to follow a sample of SSNs over the decade. This clearly would be a move towards a partial population register. Despite possible public concerns about massive databases, having a statistical population register as a goal might be a good way to rationalize and prioritize intercensal activities. The goal of a household address register, updated transactionally, seems evident already in the work that the Census Bureau is undertaking with its improving Master Address File (MAF) system.

These ideas for decennial uses of administrative records can be intertwined with suggestions regarding the Census Bureau's current survey program, as we will discuss below. In any case, much greater coordination (and positive synergy) between these two separate efforts is needed than has been true traditionally. See Alexander and Chand (1999) for the kind of effort that could really pay off.

6. CURRENT SURVEY IMPLICATIONS

The introduction of administrative records into the design and estimation of the Census Bureau's continuing surveys seems a natural step towards an ARC. Some examples of how this can be done include:

6.1 Sampling Frame Uses

The American Community Survey (ACS) might be a natural starting point for an effort to use administrative records in a multiple frame context. ACS' use of the Master Address File could be supplemented, for example, with tax return addresses and, potentially, Social Security recipient addresses – and for more than just updating addresses.

6.2 Matching Poststratified Samples

Some time ago Scheuren (1980) advocated that the CPS might be routinely matched to administrative records and that administrative controls be used as poststratifiers. The pilot for this was the 1973 Exact Match Study. The approach would be much more workable today. Work, like that of Thomsen and Zhang (1999), might form an up-to-date prototype. A related approach is found in Kennickell and Woodburn (1997).

6.3 Linking Current Survey Program to Intercensal Goals

Whether you start from an administrative frame or match back to an administrative list, each effort will provide information on coverage weaknesses in the administrative records that will make it possible for them to be better used in the intercensal period. Also, such joint operations will point out where to concentrate coverage research for 2010.

An ongoing program embedded in the current survey effort to enhance already excellent demographic methods is essential (and seems to be under consideration by Prevost and Leggieri 1999). Resistance to adjustment can be worn down by repeated and open experiments over the decade, accompanied by continuous coverage and content improvements. A goal should be set to develop an annual fully projected ARC beginning no later than 2005. Funding for a large enough sample to supplement administrative records ought to be sought, perhaps through the American Community Survey. Frankly, though, this budget strategy may require that the Census Bureau promise to make major savings in 2010 – a risky proposition but necessary psychologically and fiscally.

7. ADDITIONAL 2010 RESEARCH IMPLICATIONS

Specific suggestions for additional 2010 research are hard to make, since much will depend on how successful the Census Bureau is in making their other administrative record uses serve multiple purposes. Nonetheless, two observations may be worth highlighting in any case, since they are not mentioned above.

7.1 Tracing Sample from 2000 Census

A large tracing sample should be followed over the decade. The starting point might be the post-censal quality check sample, after matching it to administrative records and augmenting it with cases, to the extent feasible, found only in the census or only in an administrative record. The kind of administrative steps outlined above would be followed, plus the actual use of tracing methodologies in at least a subsample. Fieldwork would be necessary to sort out all the problems in "cross-footing" satisfactorily from one census to another. Most of the work would be done by matching in successive waves of administrative records (in a manner similar to that touched on in section 5.2). Again, a big issue would be privacy concerns (as set out already in subsection 3.2).

7.2 Special Censuses

To prepare for 2010, there will be a need to conduct special censuses that begin with administrative records and attempt to complete them using sampling. In structure, these would not be very different from the pretests done before every census. However, because the ARC paradigm is new, there would need to be more tests and, especially, more testing time. The first 2010 tests should be built into the 2000 Census and should continue uninterrupted through the decade. Early on, general feasibility issues need to be addressed. For example –

7.2.1 Developing a way to efficiently use an administrative amalgam of addresses and individual names as a frame, so that addresses not on the administrative list are oversampled.

7.2.2 Developing a way to efficiently handle multi-unit dwellings, since the administrative addresses usually do not have apartment numbers. (It may be that some of the sampling will have to be independent of the lists, as in the census quality check sampling, then matched-in after the fact.)

7.2.3 Developing an approach for dealing with problem populations (*e.g.*, low-income minority children) will need special attention (Medicaid data, mentioned earlier, might be of help here but this is unclear).

7.2.4 Developing a means to deal with problem locations in the 2000 Census (*e.g.*, inner city neighborhoods) may need to be looked at individually.

The notion of designing these special censuses as a rolling sample (*e.g.*, Kish 1990) would allow – say, by the end of 2005 – a way to obtain a "gold standard" for evaluating the ARC approach that evolves. Note, there is no reason that two or more methods cannot be tried simultaneously to speed up the process of testing. Indeed, it may turn out that the 2010 Census should employ multiple approaches simultaneously – including sampling. Given the diversity of circumstances that exist, multiple approaches may prove inherently better than any single approach.

8. A SUMMARY AND SOME POSSIBLE PRIORITIES

The U.S. Census Bureau's major efforts (*e.g.*, Prevost and Leggieri 1999) to research an administrative record census are deserving of applause. Even though the Census Bureau is now well underway in its ARC research, it still might be of value to reiterate key points and priorities.

8.1 Constancy of Purpose

Deming, in setting out his famous 14 points, lists "Constancy of Purpose" or, in the words of the old Negro spiritual, "Keep your eyes on the prize." With all the extra challenges of running a census in 2000, keeping focused on the future may be the hardest task facing the excellent staff assembled. Temptations to cut budgets or reassign key people must be avoided. After the decennial census, separate budgeting should be sought and the sums involved will need to be large. Thinking that the big efforts are connected with the 2000 Census and could then slack off for a while is just flat wrong. The research effort will need to grow and grow.

8.2 Environmental Scan

While the responsibility for the official census count will not change any time soon (if ever), census-taking will no longer be the monopoly it has been. Ways to integrate independent information sources will be essential to how the Census Bureau's success is measured. Levels of accountability can be predicted to increase. The Connecticut case study, discussed in section 3, is just one example.

What is crucial to see is that, ironically, central statistical agencies – including places like the Census Bureau – could well be left behind in the information age. Census Bureau market share in the information sector has been falling for decades and, *ceteris paribus*, a steeper drop is quite likely during the next ten years, given the slow pace of change inherent in a government agency. The Census Bureau's administrative record research and its continuing emphasis on being the leader in key information technologies could mitigate this trend, but probably not reverse it.

8.3 Assumptions

The privacy assumptions are the ones to worry about the most, as leaders at the Bureau, like Gates, have long been saying. Careful watching and listening are needed. The use of an advisory Institutional Review Board, not mentioned earlier, might be considered – to provide independent oversight with the public's interest in mind, especially on record linkage. Alternatively, now that there is a Census Monitoring Board, the Board might be the natural place to focus advice on handling privacy concerns and in doing priority setting.

The real concern is not that the Census Bureau will proceed rashly, but that it might be too timid. To quote Emerson, "Be bold, be bold, be not too bold." Certainly the Census Bureau should seek legislation like the Statistics Act in Canada, in order to assure the cooperation of administrative agencies in providing data for an ARC. As already noted, the development and enactment of such legislation would provide the opportunity for a public debate on ARC ideas, something that must occur before an ARC could be done.

8.4 Suggestions for 2000 Census

The inclusion of the SSN question in the quality check sample for 2000 is crucial to building the bridge from old to new. The suggestion to produce a simultaneous partial ARC estimate will help not only in testing an approach that will be needed in 2010, but also in validating both the ARC and the census itself. Regarding other administrative record research, carpe diem – seize the moment – especially the opportunity to acquire key files (an effort already well underway according to Prevost and Leggieri 1999) and to strengthen long-term partnerships that build human capital.

8.5 Intercensal Steps

Creating a greater positive synergy between the current monthly and annual survey programs and the intercensal estimates program is key too. Transforming the intercensal estimates effort to one that is transaction-based, rather than essentially cross-sectional, would be the other major priority. Integrating special census results would also be crucial.

8.6 Current Survey Steps

The opportunities for administrative record applications in the American Community Survey are excellent, if they continue to be grasped quickly enough. However, the time from inception to results for new census surveys is often too long. Censuses have cycle times that extend arguably over more than ten years. The introduction of a new frame in the current survey programs has been growing. Whatever is done after the next census, it needs to be a lot quicker than after the last.

8.7 Additional Possible Steps

Of the two suggestions in the last section, the tracing sample has the most appeal as basic research. The Canadian experience can help here, but payoffs are uncertain. Tracing would help in addressing immigration flows, both legal and illegal. Obviously, as proof of concept and to make the ARC operationally feasible, special censuses will be critical. The budgeting will have to be a lot heavier in the early years of the decade than historically has been the case for the census pilots and dress rehearsals done in the 1980's and 1990's. Planning for the continuing research after 2010 should be a priority, as well, and might begin now and be revisited at least annually.

In the March 26, 1999, issue of *Science* (Cohen 1999), there is a news item entitled "The March of Paradigms." It tracks the growing number of scientific papers that use the phrase "new paradigm" in their titles or abstracts. It goes on to state that –

Many of these claims, however, may not be quite the kind of developments science philosopher Thomas Kuhn had in mind when he made the term new paradigm famous with his paradigm-shifting 1962 book, *The Structure of Scientific Revolutions.*

Despite this caution, the change to a partial ARC does qualify as a paradigm-shifting event and should be studied from that perspective. In this connection, compliments are due to all those who have already attempted and achieved it around the world. Best wishes to the U.S. Census Bureau in their research on it now.

ACKNOWLEDGEMENTS

For the ideas in this paper there are many who deserve thanks. Some of the key people are Wendy Alvey, Bonny Bye, John Leyes and Peter Sailer. Thanks are also owed to the Associate Editor and two very helpful referees.

REFERENCES

ALVEY, W., and SCHEUREN, F. (1982). Background for an administrative records census. (With discussion by John Leyes, Statistics Canada). *Statistics of Income and Related Administrative Record Research.* Washington, DC: U.S. Department of the Treasury, Internal Revenue Service.

ALEXANDER, C., and CHAND, N. (1999). Indirect estimation with administrative records and the American Community Survey. *1999 Federal Committee on Statistical Methodology Proceedings.* Washington DC: U.S. Bureau of the Census.

BELLHOUSE, D. (1988). *Hanbook of Statistics: Sampling, A brief history of random sampling methods.* New York: North-Holland.

BISKUPICK, J. (1998). Division of representation, funds at stake in census feud. *The Washington Post.* 27 November 1998.

BLUM, O. (1999). Combining register-based and traditional census processes as a pre-defined strategy in census planning. *1999 Federal Committee on Statistical Methodology Proceedings.* Washington DC: U.S. Bureau of the Census.

BYE, B. (1997). Administrative Record Census for 2010: Design Proposal. Prepared for the U.S. Bureau of the Census. Rockville MD: Westat Inc.

BYE, B. (1998a). Race and Ethnicity Modeling with SSA Numident Data: File Development and Tabulations. Prepared for the U.S. Bureau of the Census. Rockville MD: Westat Inc.

BYE, B. (1998b). Race and Ethnicity Modeling with SSA Numident Data: Individual-level Regression Model - Version 2. Prepared for the U.S. Bureau of the Census. Rockville MD: Westat Inc.

BYE, B. (1999). Race and Ethnicity Modeling with SSA Numident Data: Two-level Regression Model. Prepared for the U.S. Bureau of the Census. Rockville MD: Westat Inc.

CARTER, R., and McCLEAN, K. (1996). Using administrative data in the Canadian census: experiences and plans. *Statistical Journal of the United Nations,* 13, 4, 375-383.

CLARK, C., and GATES, G. (1999). *Memorandum on Restricted Access Policy for Administrative Records.* U. S. Bureau of the Census, June 25, 1999.

COHEN, J. (1999). The march of paradigms. *Science,* 283, 1998-99.

CZAJKA, J.L., and WALKER, B. (1989). Combining panel and cross-sectional selection in an annual sample of tax returns. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 463-468.

CZAJKA, J.L., MORENO, L., and SCHIRM, A. (1997). On the Feasibility of Using Internal Revenue Service Records to Count the U.S. Population. Washington, DC: Mathematica Policy Research.

DEMING, W. (1986). *Out of the Crisis.* Cambridge MA: MIT Press.

EDMONSTON, B., and SCHULTZE, C. (Eds.) (1995). Modernizing the U.S. Census, Panel on Census Requirements in the Year 2000 and Beyond. Committee on National Statistics, National Research Council, Washington, DC: National Academy Press.

FELLEGI, I. (1997). Record linkage and public policy - a dynamic evolution. *Record Linkage Techniques,* 3 - 12. Arlington VA: Ernst and Young, LLP.

GATES, G. (1999). Data Mining, Panel on Privacy and Statistics in the New Millennium. Panel presentation at the Joint Statistical Meetings, Baltimore, MD.

GATES, G., and BOLTON, D. (1998). Privacy research involving expanded statistical uses of administrative records. *Proceedings of the Government and Social Statistics Section, American Statistical Association.*

HANSEN, M., and HURWITZ, W. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics,* 14, 333-62.

INTERNATIONAL STATISTICAL INSTITUTE (1994). *The Future of Statistics,* (Z. Kenessey, Ed.). ISBN: 90-73592-11-9.

JABINE, T., and SCHEUREN, F. (1987). Record linkages for statistical purpose: methodological issues. *Journal of Official Statistics,* 2, 255-277.

JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association.*

JENSEN, P. (1983). Towards a register-based statistical system - some Danish experiences. *Statistical Journal of the United Nations*, 341-365.

JOHNSON, N., and KOTZ, S. (Eds.) (1997). *Leading Personalities in Statistical Science: From the Seventeenth Century to the Present.* New York: Wiley.

KENESSEY, Z. (Ed.) (1994). *The Future of Statistics: An International Perspective.* Voorburg: International Statistical Institute.

KENNICKELL, A., and WOODBURN, L. (1997). Consistent Weight Design for the 1989, 1992, and 1995 SCF, and the Distribution of Wealth, available from the web site http://www.bog.frb.fed.us/pubs/oss/oss2/scfindex.html.

KISH, L. (1990). Rolling samples and censuses. *Survey Methodology*, 16, 1, 63-71.

KUHN, T. (1970). *The Structure of Scientific Revolutions.* Chicago IL: University of Chicago Press.

LEYES, J., and ELSL-CULKIN, J. (1994). Administrative social data in Canada: some results and some implications. *Statistics of Income: Turning Administrative Systems into Information Systems*, U.S. Internal Revenue Service: Washington, DC.

MYRSKLA, P. (1991). Census by questionnaire – census by registers and administrative records: the experience of Finland. *Journal of Official Statistics*, 7, 457-74.

NATIONAL ACADEMY OF SCIENCES (1996). Massive Data Sets: Proceedings of a Workshop Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences, National Research Council: Washington, DC.

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

PREVOST, R., and LEGGIERI, C. (1999). Expansion of administrative record uses at the census bureau; a long-range research plan. *1999 Federal Committee on Statistical Methodology Proceedings*, Washington DC: U.S. Bureau of the Census.

REDFERN, P. (1989). Population registers: Some administrative and statistical pros and cons. *Journal of the Royal Statistical Society*, Series A, 153, 1-41.

SAILER, P., and WEBER, M. (1998). The IRS population undercount: an update. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

SAILER, P., WEBER, M., and YAU, E. (1993). How well can the IRS count the population? *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

SCHAIBLE, W. (Ed.) (1996). *Indirect Estimators in U.S. General Programs.* New York: Springer-Verlag.

SCHEUREN, F. (1979). Integrated linked administrative statistical sample. LASS Working Notes, U.S. Social Security Administration.

SCHEUREN, F., OH, H.L., VOGEL, L., and YUSKAVAGE, R. (1981). Methods of Estimation for the 1973 Exact Match Study. Studies from Interagency Data Linkages. U.S. Department of Health and Human Services, Social Security Administration, Publication 13-11750.

SCHEUREN, F. (1985). Methodological issues in linkage of multiple databases. In *Record Linkage Techniques - 1985: Proceedings of the U. S. Internal Revenue Service*, 155-178.

SCHEUREN, F. (1990). Discussion of Kish (1990). *Survey Methodology*, 16, 1, 72-79.

SCHEUREN, F. (1995a). A U.S. Administrative records census. *Chance*, 8, 2, 43-45.

SCHEUREN, F. (1995b). Private lives and public policies: confidentiality and accessibility of government services. In *Journal of the American Statistical Association*, 90, 386-387. Washington, DC: National Academy Press (1993).

SCHEUREN, F. (1997). Linking health records: human rights concerns. *Record Linkage Techniques.* Washington DC: Ernst and Young, LLP.

SCHEUREN, F., and PETSKA, T. (1993). Turning administrative systems into information systems. *Journal of Official Statistics*, 9, 109-119.

SCHEUREN, F., and WINKLER, W. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

SCHEUREN, F., and WINKLER, W. (1997). Regression analysis of data files that are computer matched - Part II. *Survey Methodology*, 23, 157-165.

SINGER, E. (1999). Data Mining, Panel on Privacy and Statistics in the New Millennium. Panel presentation at the Joint Statistical Meetings, Baltimore, MD.

STATISTICS CANADA (1993-1996). *Survey of Labour and Income Dynamics: Research Papers.* Catalogue No. 75F0002MIE, 93-01, 94-03, 94-11, 95-19 and 96-12.

STATISTICS CANADA (1999). *Statistics Canada's Survey of Financial Security: Update - July 1999.* Catalogue 13F002MIE 99006.

STEFFEY, D., and BRADBURN, N. (Eds.) (1994). *Counting People in the Information Age, Panel to Evaluate Alternative Census Methods.* Committee on National Statistics, National Research Council, Washington, DC: National Academy Press.

SUKHATME, P. (1935). Contributions to the theory of the representative method. *Journal of the Royal Statistical Society Supplement*, 2, 263-68.

THOMSEN, I., and HOLMOY, A.M. K. (1998). Combining data from surveys and administrative record systems, the Norwegian experience. *International Statistical Review*, 66, 2, 201-221.

THOMSEN, I., and ZHANG, L. (1999). The effects of using administrative registers in economic short term statistics: the Norwegian Labour Force Survey as a case study. 1999 Federal Committee on Statistical Methodology Proceedings. Washington DC: U.S. Bureau of the Census.

WHITE, G., MULROW, E., and SCHEUREN, F. (1999). *Connecticut Jury Record Linkage Research.* Washington DC: Ernst and Young, LLP.

ZASLAVSKY, A.M., and WOLFGANG, G.S. (1993). Triple-system modeling of census, post-enumeration survey, and administrative list data. *Journal of Business and Economic Statistics*, 11, 279-288.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 161-173
Statistics Canada

A New Look at Confidence Intervals in Survey Sampling

V.P. GODAMBE and M.E. THOMPSON[1]

ABSTRACT

In survey sampling, as in other areas of statistics conventionally, confidence intervals for a parameter are often obtained by inverting the distribution of some approximate pivotal quantity, {(estimate – parameter)/(estimated variance)^{½}}. Alternatively, estimating function theory suggests a more direct method of constructing a pivotal quantity and hence confidence intervals. These alternative confidence intervals perform much better than the conventional ones in simulation studies.

KEY WORDS: Confidence intervals; Estimating functions; Optimality; Stratification; Survey sampling.

1. HISTORICAL INTRODUCTION

The topic of confidence intervals was first discussed in Neyman's (1934) well-known paper read before the Royal Statistical Society. The paper was on survey sampling. Yet Neyman's discussion did not arrive at the actual construction of confidence intervals for a survey sampling setup. This may have been due to the fact that at the time the distinction between the parameters of a survey population on one hand and a hypothetical population on the other was far from clearly understood (Deming 1950, Godambe 1976, Godambe 1997, Smith 1997). In hindsight, one can say that Neyman's discussion of confidence intervals related primarily to the parameters of a hypothetical population. A subsequent publication of Neyman (1937) explicitly demonstrated how confidence intervals could be obtained from a pivotal quantity, a function of observations and the parameter of interest having a fixed (known) distribution. The availability of such "pivotals" (or of approximate pivotals) for certain hypothetical populations characterized by a few scalar parameters can be easily demonstrated. On the other hand, to characterize a survey population of size N one needs a parameter of N dimensions (Basu 1958, Hájek 1959). Under this condition, in general, no nontrivial function of the observations and the parameter of interest can be exactly pivotal under the distribution induced by a probability sampling design.

Section VI of Neyman's 1934 paper is entitled Appendix. In addition to other things, the Appendix contains Note I, dealing with confidence intervals, followed by Note II, "The Markoff Method and Markoff Theorem on Least Squares". The "Theorem" mentioned here, using modern terminology, is the Gauss-Markoff theorem on unbiased minimum variance estimation. Now it is true that the "variance" of an unbiased estimator, if known, can enable one to construct an approximate confidence interval by assuming an approximate normal distribution for the estimator: (estimator – parameter)/(variance)^{½} is an approximate pivot. This however is of no avail, for the "variance" just mentioned is never known in a survey sampling situation. A common practice, as seen from publications on the subject (*e.g.*, Chaudhuri and Vos 1988), is to substitute an "estimate" for the unknown variance. Here the basic question, generally not discussed in the literature, is: which of the many estimates of the variance would provide a pivot (or approximate pivot) leading to a set of plausible confidence intervals? This problem for a hypothetical population with an underlying parametric model is resolved utilizing the observed Fisher information (Efron and Hinkley 1978). A generalization of the observed Fisher information, for a semiparametric model, provided by the theory of optimal estimating functions (Godambe 1985, Godambe and Thompson 1986) leads to an answer to the question just raised, within the context of survey sampling.

So far the topic of "confidence intervals in survey sampling" has been considered within the framework of "estimating functions" in only a couple of papers. Historically, Woodruff (1952) presented an earliest demonstration of confidence intervals for "position measures" of a survey population, utilizing estimating functions informally. This argument has been commented on in detail by Godambe (1991). The second paper is by Binder and Patak (1994). The first author, in an earlier publication (Binder 1983), made informal use of estimating functions for complex surveys. There, however, the confidence intervals presented were of more conventional type. Both the paper by Binder and Patak (1994) and the present paper are based on the theory of estimating functions. The basic difference between the two is that the latter, in an essential manner, is tied to the optimality criterion of estimating functions. This optimality criterion relates present survey populations to a semiparametric superpopulation model, albeit very flexibly. This relationship, as the present paper demonstrates, provides guidance in the choice of estimating function and the implied confidence intervals to be used for a given problem. Apart from this reference to a superpopulation model, the confidence intervals presented in this paper are design-based. Again, both papers, Binder and Patak (1994)

[1] V.P. Godambe and M.E. Thompson, Statistics and Acturial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

and the present one (section 5), discuss the important case of "nuisance parameters". However the problems treated in the two papers are different and there is no overlap in the results.

2. MODEL-ROBUST CONFIDENCE INTERVALS

We begin the discussion by talking about the estimation of a single parameter θ of a superpopulation model. Suppose that, under the model, the observations y_i for i in the sample s are independent, with

$$\varepsilon\, g_i(y_i, \theta) = 0, i \in s;$$

$$\varepsilon\, g_i^2(y_i, \theta) = \sigma_i^2, i \in s, \tag{2.1}$$

where the g_i are elementary estimating functions.

We can obtain "model-robust" approximate confidence intervals for θ by inverting

$$\left| \left\{ \sum_{i \in s} g_i(y_i, \theta) \right\} \Big/ \left\{ \sum_{i \in s} g_i^2(y_i, \theta) \right\}^{\frac{1}{2}} \right| = z \tag{2.2}$$

where z is a percentile of the $N(0, 1)$ distribution.

A very general version of (2.2) was put forward as an approximate pivotal, in a framework of stochastic processes, by one of the authors (Godambe 1985). Previous versions of (2.2), by Fisher (1925), Efron and Hinkley (1978) and Royall (1986), all used the numerator (estimate $\hat{\theta}$ – parameter θ); the denominators in all three cases were the square roots of different estimators of the variance of $(\hat{\theta} - \theta)$. For an early investigation of properties of (2.2), see Mach (1988). A later study was carried out by Vinod (1998).

There are several parts to the rationale for the use of (2.2):

(a) $\sum_{i \in s} g_i^2(y_i, \theta)$ is a model-unbiased estimator of $\mathrm{Var}(\sum_{i \in s} g_i(y_i, \theta))$, regardless of the form of $\{\sigma_i^2\}$;

(b) being analogous to the observed information, $\sum_{i \in s} g_i^2(y_i, \theta)$ can be thought of as conditionally unbiased, given important aspects of the sample structure; more specifically, $\sum_{i \in s} g_i^2(y_i, \theta)$ is the "variance" of the numerator "conditional" on the same partitioning which underlies the optimality of the numerator (Godambe 1985; Godambe and Thompson 1986);

(c) if the model is misspecified, $\sum_{i \in s} g_i^2(y_i, \theta)$ will incorporate to some extent the bias of the estimating function as well as its variability;

(d) if $y_i, i \in s$ are i.i.d. $N(\theta, \sigma^2)$ then

$$\tau = \frac{\sum_{i \in s} (y_i - \theta)}{\sqrt{\sum_{i \in s} (y_i - \theta)^2}} \tag{2.3}$$

is closer to $N(0, 1)$ than is the t-statistic, since $\mathrm{Var}(\tau) = 1$, and the kurtosis of τ is $3\text{-}6/(n+2)$; confidence intervals for θ based on inverting $|\tau| = z$ are

$$\bar{y} \pm \sqrt{\frac{n-1}{n-z^2}}\, \frac{z s_y}{\sqrt{n}} \tag{2.4}$$

where s_y is the sample standard deviation.

Suppose now that $\boldsymbol{\theta}$ is a vector-valued parameter, and that $\psi(\boldsymbol{\theta})$ is a scalar parameter of interest. Suppose that the y_i are independent under the superpopulation model, that $g_i(y_i, \boldsymbol{\theta})$ has the same dimensionality as $\boldsymbol{\theta}$, and that the unbiased estimating equation system

$$\sum_{i \in s} g_i(y_i, \boldsymbol{\theta}) = 0 \tag{2.5}$$

arises from the minimization of a scalar objective function

$$\sum_{i \in s} G_i(y_i, \boldsymbol{\theta}). \tag{2.6}$$

Estimation of $\psi(\boldsymbol{\theta})$ could proceed by "profiling", that is by finding $\tilde{\boldsymbol{\theta}}(\psi)$ which would minimize (2.6) for a fixed value $\psi(\boldsymbol{\theta}) = \psi$. Then the estimating function for ψ would be one which found $\hat{\psi}$ to minimize

$$\sum_{i \in s} G_i(y_i, \tilde{\boldsymbol{\theta}}(\psi)).$$

The vector form of the system for finding $\boldsymbol{\theta}\,(\psi)$ would be

$$\sum_{i \in s} g_i(y_i, \boldsymbol{\theta}) - \lambda \frac{\partial \psi}{\partial \boldsymbol{\theta}} = 0, \tag{2.7}$$

where λ is a (scalar) Lagrange multiplier, together with the constraint that $\psi(\boldsymbol{\theta}) = \psi$. There is a one-to-one correspondence between ψ and λ, with λ being 0 when ψ is $\hat{\psi}$. The estimate $\hat{\psi}$ will solve

$$\sum_{i \in s} g_i(y_i, \tilde{\boldsymbol{\theta}}(\psi)) = 0 \tag{2.8}$$

or a linear combination of its components.

If $\boldsymbol{a}$ is a vector with the dimension of $\boldsymbol{\theta}$, and

$$\boldsymbol{a}^{\tau} \sum_{i \in s} g_i(y_i, \boldsymbol{\theta}\,(\psi)) = 0 \tag{2.9}$$

is a (possibly suboptimal) estimating equation for $\hat{\psi}$, it seems reasonable to obtain approximate confidence intervals for ψ by inverting

$$\frac{\left| a^{\tau} \sum_{i \in s} g_i(y_i, \theta(\psi)) \right|}{\sqrt{a^{\tau} \left[\sum_{i \in s} g_i(y_i, \theta(\psi)) g_i^{\tau}(y_i, \theta(\psi)) \right] a}} = z. \qquad (2.10)$$

Remarks:

(i) The estimating equations (2.8) and (2.9) are only approximately unbiased, and their terms are only approximately independent. Thus the use of (2.10) will be more easily justified theoretically for large samples.

(ii) Even if the system (2.5) does not arise from the minimization of an objective function (2.6), the process of estimation of ψ through (2.7), (2.9) and (2.10) can still be carried out, and is still meaningful.

(iii) When $\theta(y)$ and $\psi(y)$ are finite population parameters analogous to θ and ψ, the same process can be carried through, with (2.10) replaced by

$$\frac{\left| a^{\tau} \sum_{i \in s} g_i(y_i, \theta(\psi(y))) \right|}{\sqrt{v(\chi_s, \theta(\psi(y)))}} = z, \qquad (2.11)$$

where $y = (y_1, ..., y_N), s$ is a sample, $\chi_s = \{(i, y_i): i \in s\}$, and the denominator is a suitable estimate of the standard deviation of the numerator. (See section 3 and 7 for notation and elaboration.)

Indeed, the purpose of this paper is to explore the adaptation of (2.2) and (2.10) and their rationale to the survey sampling context. For example, suppose we have a finite population of size N, that $y_i, i = 1, ..., N$ are i.i.d. $N(\theta, \sigma^2)$, and that we wish to estimate, or equivalently "predict", the finite population mean $\bar{Y} = \sum_{i=1}^{n} y_i/N$ from the observations in a sample. The same distribution theory as in (d) above establishes that

$$\tau = \frac{n^{1/2} b(\bar{y} - \bar{Y})}{\sqrt{(n-1)s_y^2 + b^2(\bar{y} - \bar{Y})^2}}, \qquad (2.12)$$

where $b = (1/n - 1/N)^{1/2}$, is approximately $N(0, 1)$. Inverting $|\tau| = z$ gives rise to "prediction" intervals for $\bar{Y}$ of form

$$\bar{y} \pm \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} s_y z \sqrt{\frac{n-1}{n - z^2}}. \qquad (2.13)$$

Under the assumed model, these will have improved prediction properties over the usual simple random sampling based confidence intervals

$$\bar{y} \pm \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} s_y z. \qquad (2.14)$$

When the sampling design is simple random sampling, a sampling unbiased estimator of the sampling variance of $n^{1/2} b(\bar{y} - \bar{Y})$ is

$$\frac{N}{N-1} \sum_{i \in s} (y_i - \bar{Y})^2. \qquad (2.15)$$

When N is large, the pivot τ of (2.12) is approximately

$$\frac{n^{1/2} b(\bar{y} - \bar{Y})}{\sqrt{\frac{N}{N-1} \sum_{i \in s} (y_i - \bar{Y})^2}},$$

which is approximately

$$\frac{\sum_{i \in s} (y_i - \bar{Y})}{\sqrt{n \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{i \in s} (y_i - \bar{Y})^2}}. \qquad (2.16)$$

Thus, in summary, the pivot of (2.12) takes a form similar to that of (2.2), with θ replaced by the finite population parameter $\bar{Y}$. The pivot of (2.12) has both a prediction interpretation and a design based (simple random sampling) justification.

3. OPTIMAL ESTIMATING FUNCTIONS FOR SURVEY POPULATIONS

Quite commonly, estimation for survey populations is design based as well as model based. Godambe and Thompson (1986) have proposed the following framework for optimal estimation of finite population quantities which correspond to superpopulation parameters.

Let θ be a superpopulation parameter and let $\varphi_1, ..., \varphi_N$ be independent elementary estimating functions such that $\varepsilon \varphi_i(y_i, \theta) = 0$ for $i = 1, ..., N$. Let $y = (y_1, ..., y_N)$ be the population vector of responses, and let $\theta(y)$ be the finite population parameter which is the solution in θ of

$$\sum_{i=1}^{N} \varphi_i(y_i, \theta) = 0. \qquad (3.1)$$

We think of θ and $\theta(y)$ as being associated parameters, one of the superpopulation and one of the finite population. We take them to be real for simplicity.

Suppose $p = \{p(s), s \in S\}$ is a sampling design, or probability function on S, the collection of samples or subsets s of $\{1, ..., N\}$. The sampling design p induces a distribution on the outcome $\chi_s = \{(i, y_i): i \in s\}$. Let E_p denote expectation under the sampling design. A sample estimating function is a function $g_s(\chi_s, \theta)$, and an estimating function strategy (g, p) is taken to be unbiased if

$$E_p g_s(\chi_s, \theta) = \sum_{i=1}^{N} \varphi_i(y_i, \theta) \qquad (3.2)$$

for all y, θ. Point estimation for $\theta(y)$ proceeds by finding the solution $\hat{\theta}_s$ of

$$g_s(\chi_s, \theta) = 0. \tag{3.3}$$

Among unbiased strategies (g, p) (p fixed), strategy (g^*, p) may be taken to be optimal if

$$\varepsilon E_p(g_s^*(\chi_s, \theta) - \sum_{i=1}^{N} \varphi_i(y_i, \theta))^2 \tag{3.4}$$

is minimal. It was shown by Godambe and Thompson (1986) that the optimal estimating function is given by

$$g_s^*(\chi_s, \theta) = \sum_{i \in s} g_i^*(y_i, \theta) \tag{3.5}$$

where $g_i^*(y_i, \theta) = \varphi_i(y_i, \theta)/\pi_i$ and π_i is the probability (under p) that i is included in the sample. Note that the optimality of g_s^* in this sense is independent of the variance structure $\{\varepsilon\varphi_i^2(y_i, \theta)\}$. The unbiasedness constraint (3.2) is a very strong one, but also an important one in the survey sampling context.

Suppose now that we have a sampling estimating function $g_s(\chi_s, \theta) = \sum_{i \in s} g_i(y_i, \theta)$ which satisfies (3.2), and which may or may not be optimal for point estimation of $\theta(y)$ under criterion (3.4). Because $g_s(\chi_s, \theta)$ is also an estimating function for θ, the inversion of (2.2):

$$\left| \frac{\sum_{i \in s} g_i(y_i, \theta)}{\sqrt{\sum_{i \in s} g_i^2(y_i, \theta)}} \right| = z$$

should provide confidence intervals for θ with good coverage, under the superpopulation model.

When $\theta(y)$ is the object of inference, it is tempting again to take a prediction approach, and to consider inverting

$$\left| \frac{\sum_{i \in s} g_i(y_i, \theta(y))}{\sqrt{v_m(\chi_s, \theta(y))}} \right| = z \tag{3.6}$$

where

$$\varepsilon v_m(\chi_s, \theta(y)) = \varepsilon\left(\sum_{i \in s} g_i(y_i, \theta(y))\right)^2. \tag{3.7}$$

Alternatively, we may take a design-based, inverse testing approach, and consider inverting

$$\left| \frac{\sum_{i \in s} g_i(y_i, \theta(y))}{\sqrt{v_p(\chi_s, \theta(y))}} \right| = z \tag{3.8}$$

where

$$E_p v_p(\chi_s, \theta) = E_p\left(\sum_{i \in s} g_i(y_i, \theta) - \sum_{i \in s}^{N} \varphi_i(y_i, \theta)\right)^2. \tag{3.9}$$

For a well chosen design which corresponds with appropriate elements in the model, these two approaches should give results which are close to one another; and the intervals for $\theta(y)$ will be different from the intervals given by (3.6), by an amount which takes into account the finiteness of the population.

Similar considerations apply for a multidimensional parameter $\theta(y)$. However, in the next section we will illustrate the general approach with a one dimensional parameter, in the somewhat artificial situation of strata with a common mean.

4. STRATIFIED SIMPLE RANDOM SAMPLING

We follow the usual notation. The labelled population of N individuals (units) is denoted by $\mathcal{P} = \{i : i = 1, \ldots, N\}$. The population $\mathcal{P}$ is divided into k nonoverlapping strata $\mathcal{P}_j$ of sizes $N_j, j = 1, \ldots, k$. A variate of study defined for the population $\mathcal{P}$ is y, assumed to be scalar for simplicity. For the individual $i, y = y_i, i = 1, \ldots, N$. The population vector $\boldsymbol{y} = (y_1, \ldots, y_N)$. To obtain an estimate for the population mean $\bar{Y} = \sum_1^N y_i/N$ a sample s of size n is drawn from $\mathcal{P}$, with a stratified simple random sampling without replacement design. The samples from different strata $\mathcal{P}_j$ are denoted by $s_j, |s_j| = n_j, j = 1, \ldots, k; \sum n_j = n$. Further, $\bar{Y}_j$ and $\bar{y}_j$ denote the means of y in stratum $\mathcal{P}_j$ and sample s_j respectively, $j = 1, \ldots, k$. Thus

$$\bar{Y} = \sum_{j=1}^{k} N_j \bar{Y}_j / N. \tag{4.1}$$

Analogously we define

$$\bar{y} = \sum_{j=1}^{k} N_j \bar{y}_j / N. \tag{4.2}$$

If the components of the population vector $\boldsymbol{y} = (y_1, \ldots, y_N)$ are assumed to have been drawn independently from a superpopulation with a common mean $\varepsilon(y_i) = \theta, i = 1, \ldots, N$ but possibly different variances $\varepsilon(y_i - \theta)^2, i = 1, \ldots, N$, and if we regard $\bar{Y}$ as the solution of $\sum_{i=1}^{N} \varphi_i(y_i, \theta) = 0$, where $\varphi_i(y_i, \theta) = y_i - \theta$, then the optimal estimating function for estimating the population mean $\bar{Y}$ in (4.1) is given by

$$g = \sum_{j=1}^{k} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in s_j} (y_i - \bar{Y}) \tag{4.3}$$

(Godambe and Thompson 1986; Godambe 1995). This g has the form of $g_s^*(\chi_s, \theta(y))$ where g_s^* is given in (3.5). Thus the optimal estimate of $\bar{Y}$ is given by $\bar{y}$, the solution in $\bar{Y}$ of the equation $g = 0$. We will call superpopulation models defined only by first few moments, such as the model just mentioned defined by $\varepsilon(y_i) = \theta, i = 1, \ldots, N$, semiparametric, in contrast to the fully parametric models specified by the density functions.

The "optimum estimating function" for a semiparametric model has many statistically important properties in common with the "score function" for a parametric model. Hence in the semiparametric model the optimum estimating function is called a quasi-score function. (Godambe 1985; Godambe and Heyde 1987; Godambe and Thompson 1989). For a parametric model, one can construct confidence intervals using the Fisher information (defined as the variance of the score function), or its natural estimate the observed Fisher information. Similarly in case of a semiparametric model the confidence intervals can be obtained from the quasi-score function *i.e.*, the optimum estimating function, and its estimated variance. In the survey context, although the optimality criterion is tied to the superpopulation model $\varepsilon(y_i) = \theta$, the properties of the confidence intervals given below are mostly if not entirely design-based.

The design-induced variance of the optimum estimating function g in (4.3) is given by

$$V(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{N_j - 1}\sum_{i \in \mathcal{P}_j} (y_i - \bar{Y}_j)^2. \quad (4.4)$$

Further, since our parameter of interest is $\bar{Y}$, the unobserved y_i's and the stratum means $\bar{Y}_j$ in (4.4) are nuisance parameters. The superpopulation model underlying the estimating function g in (4.3), namely $\varepsilon(y_i) = \theta, i = 1, \ldots, N$, suggests that for large strata sizes N_j we may ignore the differences $\bar{Y}_j = \bar{Y}$, and replace $\bar{Y}_j$ by $\bar{Y}, j = 1, \ldots, k$ in (4.4). (Models for which the differences $\bar{Y}_j = \bar{Y}$ cannot be ignored are discussed in section 7.) With this replacement, an estimate of the variance $V(g)$ in (4.4) can be given by

$$\hat{V}(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{N_j}{(N_j - 1)}\frac{1}{n_j}\sum_{i \in s_j} (y_i - \bar{Y})^2. \quad (4.5)$$

$$= \left\{\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{(n_j - 1)}\sum_{i \in s_j} (y_i - \bar{y}_j)^2\right\} + R = \hat{V}_0 + R \quad (4.6)$$

where $\bar{y}_j$ is the mean of the sample $s_j, j = 1, \ldots, k$ as in (4.2). In the right hand side of equation (4.6) the first term is $O(1/n_j)$ while the second term R is $O(1/n_j^2)$. Hence for large samples, ignoring the term R, $\hat{V}$ in (4.5) reduces to the conventional estimate $\hat{V}_0$. This leads to the conventional confidence intervals for $\bar{Y}$ based on the inversion of the distribution of $\{g/(\hat{V}_0)^{1/2}\}$. However when the sample sizes $n_j, j = 1, \ldots, k$ are not very large, estimating function theory suggests confidence intervals for $\bar{Y}$ based on the $N(0, 1)$ asymptotic distribution of $\{g/(\hat{V})^{1/2}\}$.

For a stratified simple random sampling design, let the estimating function g in (4.3) be written as

$$g = \sum_{i \in s} g_i, \quad (4.7)$$

where as before s denotes the sample (of individuals drawn from all strata). Then, for large n_j and N_j, ignoring the finite stratum correction, and replacing $(n_j - 1)$ by $n_j, j = 1, \ldots, k$ in (4.5), we have

$$\hat{V}(g) \approx \hat{V}_a(g) = \sum_{i \in s} g_i^2 = \sum_{j=1}^{k} \frac{N_j^2}{N^2}\frac{1}{n_j^2}\sum_{i \in s_j} (y_i - \bar{Y})^2. \quad (4.8)$$

It is easy to see that in view of simple random sampling from each stratum (again for reasonably large n_j and $N_j, j = 1, \ldots, k$) the sampling distribution and the superpopulation distribution of the quantity $\{g/(\hat{V}_a)^{1/2}\}$ would tend to be the same, namely approximately $N(0, 1)$. We have already identified the optimum estimating function g with the quasi-score function. Further, just as for a parametric model the inversion of the distribution of {score function/(observed Fisher information)$^{1/2}$} provides asymptotically the shortest confidence intervals, for the semiparametric model $\{g/(\hat{V}_a)^{1/2}\}$ provides asymptotically shortest confidence intervals (Wilks 1938; Godambe and Heyde 1987).

The above analysis can be easily extended to include a covariate. Suppose for the population $\mathcal{P} = \{i : i = 1, .., N\}$ in addition to the variate y under study is defined a covariate x, again for simplicity assumed to be a scalar like y. For the individual $i, x = x_i$ is known, $i = 1, \ldots, N$. The superpopulation model $\varepsilon(y_i - \theta) = 0$ underlying the foregoing discussion is now extended to $\varepsilon(y_i - \theta x_i) = 0, i = 1, \ldots, N$. Along the lines of (4.1) and (4.2) we define

$$\bar{X} = \sum_{j=1}^{k} N_j \bar{X}_j / N \quad (4.9)$$

and

$$\bar{x} = \sum_{j=1}^{k} N_j \bar{x}_j / N \quad (4.10)$$

where $\bar{X}_j$ and $\bar{x}_j$ are the means of x in stratum $\mathcal{P}_j$ and sample s_j respectively. For estimating $\bar{Y}/\bar{X}$, the solution of $\sum_{i=1}^{N} (y_i - \theta x_i) = 0$, the optimal estimating function g in (4.3) is now replaced by

$$g = \sum_{j=1}^{k} \frac{N_j}{N} \cdot \frac{1}{n_j}\sum_{i \in s_j}\left(y_i - \frac{\bar{Y}}{\bar{X}} x_i\right). \quad (4.11)$$

As before, the solution of the equation $g = 0$ provides the optimal (or approximately optimal) estimate for $\bar{Y}$. Again, the superpopulation model $\varepsilon(y_i - \theta x_i) = 0, i = 1, \ldots, N$, suggests taking

$$\frac{\bar{Y}_j}{\bar{X}_j} = \frac{\bar{Y}}{\bar{X}}$$

when the stratum sizes N_j are large, and ignoring the differences

$$\bar{Y}_j - \frac{\bar{Y}}{\bar{X}}\bar{X}_j,\ j = 1, ..., k.$$

This leads to the following estimator of the variance of g in (4.11):

$$\hat{V}(g) =$$

$$\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) \frac{N_j}{(N_j - 1)} \frac{1}{n_j} \sum_{i \in s_j} \left(y_i - \frac{\bar{Y}}{\bar{X}} x_i\right)^2 \cdot \quad (4.12)$$

(Models for which the differences $\bar{Y}_j - (\bar{Y}/\bar{X})\bar{X}_j$ cannot be ignored are discussed in section 7.) Note that (4.12) reduces to (4.6) if x_i = constant, $i = 1, ..., N$. Again, according to estimating function theory, the confidence intervals for $\bar{Y}$ can be obtained by inversion of the sampling distribution of the (approximate) pivot $g/\{\hat{V}(g)\}^{1/2}$; the distribution asymptotically is $N(0, 1)$.

5. STRATIFIED CLUSTER SAMPLING

In this section we assume the whole population of individuals (units) is divided as before into a number of nonoverlapping strata. But now, in addition, each stratum is divided into a number of nonoverlapping clusters of individuals. The first stage sampling consists of drawing from each stratum a small number of clusters with simple random sampling. Next, from each selected cluster a sample of (ultimate) individuals is drawn, possibly with a multistage "sampling design". This sampling design is "specific" to the "cluster" and does not depend on what other clusters have been selected at the first stage of selection.

To accommodate the above situation in our framework we use the following extension of the previous notation. As before i denotes the "individual". A "cluster" is denoted by c. The elements of strata $\mathcal{P}_j$, $j = 1, ..., k$ are now clusters c; the stratum $\mathcal{P}_j$ consists of N_j clusters, $j = 1, ..., k$. A sample of individuals from cluster c is denoted by s^c and the set of clusters selected from the stratum $\mathcal{P}_j$ is denoted by s_j, with $|s_j| = n_j$ and $|\mathcal{P}_j| = N_j, j = 1, ..., k$. Otherwise we use the same notation as before. Again, the superpopulation model as before is $\varepsilon(y_i - \theta x_i) = 0$ for all individuals i in the population.

Now suppose the sampling design for the cluster c is such that (once the cluster is selected) the probability of including an individual $i \in c$ in the sample is π_i'. Hence if the cluster $c \in P_j$, the unconditional inclusion probability of i is $\pi_i'(n_j/N_j)$. Thus if the population means of y and x are denoted by $\bar{Y}$ and $\bar{X}$ respectively, the optimal estimating function for $\bar{Y}$ or $(\bar{Y}/\bar{X})$, with respect to the superpopulation model just mentioned, is given by replacing g in (4.11) by

$$g = \sum_{j=1}^{k} \frac{N_j}{n_j} \sum_{c \in s_j} \sum_{i \in s^c} \left\{ \frac{y_i - \left(\frac{\bar{Y}}{\bar{X}}\right) x_i}{\pi_i'} \right\}. \quad (5.1)$$

Further, if $\bar{Y}_c$ and $\bar{X}_c$ denote the cluster means of y and x respectively the optimal estimating function (Godambe 1995) for estimating $\bar{Y}_c$ or $(\bar{Y}_c/\bar{X}_c)$ is obtained from (5.1) as

$$g_c = \sum_{i \in s^c} \left\{ \frac{\left(y_i - \frac{\bar{Y}_c}{\bar{X}_c} x_i\right)}{\pi_i'} \right\}. \quad (5.2)$$

Now we assume that for each cluster c, the sampling design is calibrated; that is for each sample s^c of non zero selection probability,

$$\sum_{i \in s^c} \frac{x_i}{\pi_i'} = X_c,$$

where X_c is the cluster total of x. For such calibrated sampling designs, if $\hat{Y}_c$ denotes the corresponding estimate of Y_c, the cluster total of y, then from (5.1) we have

$$g = \frac{1}{N} \sum_{j=1}^{k} \frac{N_j}{n_j} \sum_{c \in s_j} \left\{ \hat{Y}_c - \left(\frac{\bar{Y}}{\bar{X}}\right) X_c \right\}. \quad (5.3)$$

With fairly straightforward algebra it can be shown that the variance of g in (5.3) satisfies

$$V(g) = E\left\{ \frac{1}{N^2} \sum_{j=1}^{k} \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left(\hat{Y}_c - \frac{\bar{Y}}{\bar{X}} X_c \right)^2 \right\} + O\left(\frac{1}{N}\right).$$

(See Appendix.) Hence if all strata sizes N_j are large enough we have

$$V(g) \approx E\left\{ \frac{1}{N^2} \sum_{j=1}^{k} \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left(\hat{Y}_c - \frac{\bar{Y}}{\bar{X}} X_c \right)^2 \right\}. \quad (5.4)$$

A natural estimate of the variance $V(g)$ in (5.4) is given by

$$\hat{V}(g) = \frac{1}{N^2} \sum_{j=1}^{k} \frac{N_j^2}{n_j^2} \sum_{c \in s_j} \left(\hat{Y}_c - \frac{\bar{Y}}{\bar{X}} X_c \right)^2 ; \quad (5.5)$$

the confidence intervals for $(\bar{Y}/\bar{X})$ or $\bar{Y}$ can be obtained as before, by inverting the sampling distribution of the approximate pivot $g/\sqrt{\{\hat{V}(g)\}}$; asymptotically the distribution is $N(0, 1)$.

The confidence intervals discussed above do not require the knowledge of the sampling design for any cluster, provided at the cluster level the estimates $\hat{Y}_c$ of Y_c are available for $c \in s_j, j = 1, ..., k$. These confidence intervals, though valid, cannot be expected to be as efficient as the ones based on the entire data, if and when available.

It is important to distinguish the estimates $\hat{V}(g)$ in (4.5), (4.12) and (5.5) (of the variances $V(g)$ of the estimating function g) from conventional estimates of estimator variances. The former generally in an essential way contain the parameter of interest Y or $\bar{Y}$. The latter by definition must be free of the parameter. We might conjecture that the distribution of $g/\sqrt{\{\hat{V}(g)\}}$ would generally tend to its limit faster than the corresponding distribution of

$$(\hat{\bar{Y}} - \bar{Y}) / \{\text{estimate of the variance of } \hat{\bar{Y}}\}^{1/2}. \qquad (5.6)$$

For unlike the {estimate of the variance of $\hat{\bar{Y}}$} in (5.6), $\hat{V}(g)$ would be a sum of independently distributed random variates, and would be stabler.

The estimate $\hat{V}(g)$ in (5.5) depends on the sample variates only through the estimates of the cluster totals or means; a property also shared by the traditional estimate of the variance of $\hat{\bar{Y}}$ in (5.6). In connection with the latter, early references can be traced back to Mahalanobis' interpenetrating samples in the thirties, while more recent examples are Särndal, Swensson and Wretman (1992), Yung and Rao (1996).

6. BOOTSTRAP VARIANCE ESTIMATION

In this section we present bootstrap versions of the estimates of the variance $\hat{V}(g)$ given in (4.5), (4.12) and (5.5). We illustrate the method in the case of (4.12) in some detail; the estimates (4.5) and (5.5) could be obtained as special cases.

Our bootstrap method is different from the usual in the sense that we obtain the bootstrap variance of the estimating function g in (4.11), holding the parameter value $(\bar{Y}/\bar{X})$ in it fixed. As before our data consists of (y_i, x_i): $i \in s_j$, $j = 1, \ldots, k$. The stratified resampling is done as follows. A number n_j of draws are made with replacement from $(y_i, x_i): i \in s_j, j = 1, \ldots, k$. If q denotes a generic draw, a generic bootstrap value g_b of the estimating function g is given by

$$g_b = \sum_{j=1}^{k} \frac{N_j}{N} \frac{1}{n_j} \sum_{q=1}^{n_j} \left(y_q - \frac{\bar{Y}}{\bar{X}} x_q \right). \qquad (6.1)$$

Denoting by E_B and V_B the bootstrap expectation and variance respectively, we have

$$E_B(g_b) = g.$$

And

$$V_B(g_b) = E_B(g_b^2) - \{E_B(g_b)\}^2 = E_B(g_b^2) - g^2. \qquad (6.2)$$

In (6.2),

$$E_B(g_b^2) = A + B + C \qquad (6.3)$$

where

$$A = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{q=1}^{n_j} E_B \left(y_q - \frac{\bar{Y}}{\bar{X}} x_q \right)^2$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{i \in s_j} \left(y_i - \frac{\bar{Y}}{\bar{X}} x_i \right)^2.$$

$$B = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{\substack{q \neq q' \\ q, q' = 1 \\ \text{stratum } j}}^{n_j} E_B \left(y_q - \frac{\bar{Y}}{\bar{X}} x_q \right) \left(y_q - \frac{\bar{Y}}{\bar{X}} x_{q'} \right), q', q$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} n_j (n_j - 1) \left(\bar{y}_j - \frac{\bar{Y}}{\bar{X}} \bar{x}_j \right)^2.$$

$$C = \sum_{\substack{j \neq j' \\ j, j' = 1}}^{k} \frac{N_j N_j'}{N^2} \frac{1}{n_j n_j'} \times$$

$$\sum_{\substack{q=1 \\ \text{stratum } j}}^{n_j} \sum_{\substack{q'=1 \\ \text{stratum } j'}}^{n_{j'}} E_B \left(y_q - \frac{\bar{Y}}{\bar{X}} x_q \right) \left(y_{q'} - \frac{\bar{Y}}{\bar{X}} x_{q'} \right),$$

$$= g^2 - \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\bar{y}_j - \frac{\bar{Y}}{\bar{X}} \bar{x}_j \right)^2.$$

We have, from the above equalities,

$$(B + C) = g^2 - \sum_{j=1}^{k} \frac{N_j^2}{n_j^2} \frac{1}{n_j} \left(\bar{y}_j - \frac{\bar{Y}}{\bar{X}} \bar{x}_j \right)^2.$$

Now because of the assumption that the variates y_i are drawn from a superpopulation satisfying $\varepsilon(y_i - \theta x_i) = 0$, $i = 1, \ldots, N$, in the above expression for $(B + C)$, $\bar{y}_j - (\bar{Y}/\bar{X})\, \bar{x}_j \simeq 0(1/\sqrt{n_j})$, $j = 1, \ldots, k$. Therefore in (6.3), for large n_j, $j = 1, \ldots, k$,

$$E_B(g_b^2) \simeq A + g^2.$$

That is in (6.2)

$$V_B(g_b) \simeq A = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \frac{1}{n_j^2} \sum_{i \in s_j} \left(y_i - \frac{\bar{Y}}{\bar{X}} x_i \right)^2 \simeq \hat{V}(g) \qquad (6.4)$$

in (4.12).

The variance estimate $\hat{V}_a(g)$ in (4.8) is obtained as a special case of (6.4) when $x_i = 1, i = 1, \ldots, N$. Similarly the variance estimate $\hat{V}(g)$ in (5.5) is obtained by replacing in (6.4) individual "i" by a cluster "c" and correspondingly replacing y_i and x_i by $\hat{Y}_c$ and X_c respectively.

7. STRATA WITH DIFFERING MEANS

The optimality of the estimating functions g in (4.3), (4.11), and (5.3) depends in an essential manner on the superpopulation condition $\varepsilon(y_i - \theta x_i) = 0, i = 1,...,N$. Given that it is the finite population parameter that is of interest, the optimality of g is not affected by the superpopulation variances $\varepsilon(y_i - \theta x_i)^2, i = 1,...,N$(Godambe 1995).

In the case where $x_i \equiv 1$ it is interesting to note that the optimality of the estimating function

$$g = \sum_{j=1}^{k} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in s} (y_i - \bar{Y})$$

in (4.3) continues to hold even when the superpopulation model $\varepsilon(y_i - \theta) = 0, i \in \mathcal{P}$ is replaced by the extended model $\varepsilon(y_i - \theta_j) = 0, i \in \mathcal{P}_j, j = 1, ..., k$. That is, now $\varepsilon\, y_i$ is allowed to vary from stratum to stratum (Godambe 1995). The optimality continues to hold because $\sum_{j=1}^{k} (N_j/N)(\theta_j - \varepsilon \bar{Y}) = 0$. However, now the variance of g cannot be approximated by replacing the stratum mean $\bar{Y}_j$ by the population mean $\bar{Y}$ in (4.4). The earlier approximation and the subsequent estimate $\hat{V}(g)$ in (4.6) were based on the assumption that (for large strata) the differences $\bar{Y}_j - \theta$ or $\bar{Y}_j - \bar{Y}, j = 1, ..., k$ could be ignored. With θ replaced in the stratum $\mathcal{P}_j$ by θ_j, the terms $\bar{Y}_j - \bar{Y}$ are no longer ignorable, $j = 1, ..., k$. Here we note that the practice of stratifying the population so as to make each stratum "internally" homogeneous tends to enlarge the differences $\bar{Y}_j - \bar{Y}, j = 1, ..., k$. The title of this section is intended to reflect this situation.

To obtain an estimate $\hat{V}(g)$, under the extended model $\varepsilon(y_i - \theta_j) = 0$, we note that

$$g = \sum_{j=1}^{k} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in s} (y_i - \bar{Y}_j)$$

and set out to estimate the nuisance parameters or $\bar{Y}_j, j = 1, ..., k$ holding $\bar{Y}$ fixed. Note that this problem of estimation is entirely different, conceptually and also mathematically, from that of the estimation of the variance of $\bar{y}$ in (4.4). The procedure is the one established in section 2, where $\psi(y)$ is $\bar{Y}$ and $\theta(y)$ consists of $\bar{Y}_j, j = 1,...,k$.

The problem of estimating $\bar{Y}_j, j = 1, ..., k$, subject to holding the population mean $\bar{Y}$ fixed, can be solved following the usual Lagrange multiplier technique. We make the working assumption that the model variances, namely $\varepsilon(y_i - \theta_j)^2 = \sigma_i^2$, are constant ($\sigma^2$) for all $i \in \mathcal{P}$. For variations of $\bar{Y}_j, j = 1, ..., k$, find a critical point of the function

$$\varphi = \sum_{j=1}^{k} \sum_{i \in s_j} (y_i - \bar{Y}_j)^2 - \lambda \left\{ \left(\sum_{j=1}^{k} \frac{N_j \bar{Y}_j}{N} \right) - \bar{Y} \right\}, \qquad (7.1)$$

where λ is the Lagrange multiplier. This technique of estimation has intuitive appeal even without reference to the superpopulation model just mentioned. It is easy to check that (7.1) is minimized for the estimate $\hat{\bar{Y}}_j$ of $\bar{Y}_j$ where

$$\hat{\bar{Y}}_j = \bar{y}_j - \frac{N_j/(n_j N)}{\sum_{j=1}^{k} [N_j^2/(n_j N^2)]} (\bar{y} - \bar{Y}), j = 1, ..., k, \qquad (7.2)$$

n_j as before being the sample sizes from stratum $j, j = 1, ..., k$. Note that when strata sizes N_j and sample sizes n_j are "proportional", that is $(n_j/N_j) = (n/N), j = 1, ..., k$, the equations (7.2) reduce to

$$\hat{\bar{Y}}_j = \bar{y}_j - (\bar{y} - \bar{Y}), j = 1, ..., k. \qquad (7.3)$$

This simple relationship can also be used when strata sizes and sample sizes are not exactly proportional but are only approximately so. Note, with reference to (7.3), that the estimating function $(\bar{Y}_j - \bar{y}_j) - (\bar{Y} - \bar{y})$ is design-unbiased.

The above discussion also suggests estimation of the stratum means $\bar{Y}_j$ when there is a non-constant covariate x. The superpopulation model underlying the estimating function g in (4.11), as noted before, was $\varepsilon(y_i - \theta x_i) = 0$, for all individuals $i \in \mathcal{P}$, with a common parameter θ. Suppose this model is to be replaced by a more flexible and realistic model where the parameter θ is allowed to vary from stratum to stratum. That is now $\varepsilon(y_i - \theta_j x_i) = 0$, $i \in \mathcal{P}_j$, $\mathcal{P}_j$ as before denoting the j^{th} stratum, $j = 1, ..., k$. As in (4), the stratum means $\bar{Y}_j$ enter the variance $V(g)$ of the estimating function g in (4.11):

$$V(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right)$$
$$\frac{1}{N_j - 1} \sum_{i \in \mathcal{P}_j} \left\{ (y_i - \bar{Y}_j) - \frac{\bar{Y}}{\bar{X}} (x_i - \bar{X}_j) \right\}^2, \qquad (7.4)$$

$\bar{X}_j, j = 1, ..., k$ as before denoting the stratum means of x's. The estimate $\hat{V}(g)$ in (4.12) was obtained by ignoring the terms $\bar{Y}_j = \bar{Y}/\bar{X}\ \bar{X}_j$ assuming large stratum sizes N_j and the superpopulation model, $\varepsilon(y_i - \theta x_i) = 0$ for all individuals $i \in \mathcal{P}$, with a "common" parameter θ. With the new, more flexible model $\varepsilon(y_i - \theta_j x_i) = 0, i \in \mathcal{P}_j, j = 1, ..., k$, the terms $\bar{Y}_j = \bar{Y}/\bar{X}\ \bar{X}_j$ cannot be ignored any longer. The appropriate estimate, namely $\hat{V}(g)$, of the variance $V(g)$ in (7.4) is given by

$$\hat{V}(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \frac{N_j}{N_j - 1}$$
$$\frac{1}{n_j} \sum_{i \in s_j} \left\{ (y_i - \bar{Y}_j) - \frac{\bar{Y}}{\bar{X}} (x_i - \bar{X}_j) \right\}^2. \qquad (7.5)$$

However now the usual confidence intervals for $(\bar{Y}/\bar{X})$ obtained by inverting the distribution of the approximate pivot $\{g/\sqrt{\hat{V}(g)}\}$ contain nuisance parameters $\bar{Y}_j$, $j = 1, ..., k$, assuming the covariate stratum means $\bar{X}_j, j = 1, ..., k$, are known.

As before we have to estimate the nuisance parameters, namely the stratum means $\bar{Y}_j$, $j = 1, ..., k$, holding the population mean $\bar{Y}$ fixed. Note that the underlying superpopulation model specifies $\varepsilon(y_i - \theta_j x_i) = 0, i \in \mathcal{P}_j, j = 1, ..., k$. Further denoting the superpopulation variances $\varepsilon(y_i - \theta_j x_i)^2 = \sigma_i^2, i \in \mathcal{P}_j$ and assuming as before, the strata sizes $|\mathcal{P}_j| = N_j, j = 1, ..., k$ to be large, we replace the function φ in (7.1) by

$$\psi = \sum_{j=1}^{k} \sum_{i \in s_j} (\sigma_i^2)^{-1} \left(y_i - \frac{\bar{Y}_j}{\bar{X}_j} x_i \right)^2 -$$

$$\lambda \left\{ \left(\sum_{j=1}^{k} \frac{\bar{Y}_j}{\bar{X}_j} N_j \bar{X}_j \right) - N\bar{Y} \right\}, \qquad (7.6)$$

λ as before being the Lagrange multiplier. In formulating ψ in (7.6) we make the working assumption that for the superpopulation model with $\varepsilon(y_i - \theta_j x_i) = 0$, the variance functions $\varepsilon(y_i - \theta_j x_i)^2 = \sigma_i^2 = \sigma^2 x_i, i \in \mathcal{P}$. That is, σ_i^2 is proportional to the covariate value x_i, $i \in \mathcal{P}$. As stated in the beginning of this section, the working assumption just mentioned is primarily for simplicity and is of no important statistical consequence (Godambe 1995). It is easy to check that the values (estimates) $\hat{Y}_j$ of $\bar{Y}_j$ which optimize in (7.6) are given by

$$\frac{\bar{y}_j}{\bar{x}_j} - \frac{\hat{\bar{Y}}_j}{\bar{X}_j} =$$

$$\left[\left\{ \left(\sum_{j=1}^{k} N_j \bar{X}_j \frac{\bar{y}_j}{\bar{x}_j} \right) - N\bar{Y} \right\} \Big/ \left\{ \sum_{j=1}^{k} \frac{(N_j \bar{X}_j)^2}{2n_j \bar{x}_j} \right\} \right] \frac{N_j \bar{X}_j}{2n_j \bar{x}_j}, \qquad (7.7)$$

$j = 1, ..., k$.

Now for the estimating function g in (4.3), the variance $V(g)$ is given by (4.4). Further if $\hat{V}_1(g)$ is the estimate of $V(g)$ based on the estimates $\hat{Y}_j$ of $\bar{Y}_j, j = 1, ..., k$ given by (7.3), then in analogy with (4.5) (but taking into account that since $\bar{Y}_j$ is estimated, the sum of squares has fewer than n_j degrees of freedom)

$$\hat{V}_1(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \cdot \frac{1}{n_j - 1}$$

$$\sum_{i \in s_j} \{ y_i - (\bar{y}_j - \bar{y} + \bar{Y}) \}^2. \qquad (7.8)$$

The confidence intervals for $\bar{Y}$ are obtained by inverting the distribution of the approximate pivot $[g / \{ \hat{V}_1(g) \}^{1/2}]$; asymptotically,

$$g \Big/ \left\{ \hat{V}_1(g) \right\}^{1/2} \sim N(0, 1). \qquad (7.9)$$

Similarly in case of a covariate, for the estimating function g in (4.11), if $\hat{V}_2(g)$ denotes the estimate of the variance $V(g)$ in (7.4) based on the estimates $\hat{Y}_j$ given by (7.7), then

$$\hat{V}_2(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{(n_j - 1)}$$

$$\sum_{i \in s_j} \left\{ \left(y_i - \frac{\bar{y}_j}{\bar{x}_j} \bar{X}_j + A_j \right) - \frac{\bar{Y}}{\bar{X}} (x_i - \bar{X}_j) \right\}^2$$

where

$$A_j = w_j' N(\bar{y}_R - \bar{Y})/N_j,$$

$$w_j' = [N_j^2 \bar{X}_j^2 / n_j \bar{x}_j] \Big/ \sum_{j=1}^{k} [N_j^2 \bar{X}_j^2 / n_j \bar{x}_j],$$

$$\bar{y}_R = \sum_{j=1}^{k} (N_j/N) \bar{X}_j \bar{y}_j / \bar{x}_j.$$

However, a less complicated form which is still a function of $\bar{Y}_j$ only through $\bar{Y}/\bar{X}$ is

$$\hat{V}_2(g) = \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\frac{1}{n_j} - \frac{1}{N_j} \right) \frac{1}{(n_j - 1)}$$

$$\sum_{i \in s_j} \left\{ \left(y_i - \frac{\bar{y}_j}{\bar{x}_j} \bar{X}_j \right) - \frac{\bar{Y}}{\bar{X}} (x_i - \bar{X}_j) \right\}^2. \qquad (7.10)$$

Again, the confidence intervals for $\bar{Y}$ are based on the inversion of the distribution of $[g / \{ V_2(g) \}^{1/2}]$; asymptotically

$$g \Big/ \left\{ \hat{V}_2(g) \right\}^{1/2} \sim N(0, 1). \qquad (7.11)$$

8. EMPIRICAL PROPERTIES

In the preceding section we have provided construction of confidence intervals when the superpopulation model $\varepsilon(y_i - \theta) = 0$ or $\varepsilon(y_i - \theta x_i) = 0$, with a "common" value of θ for all individuals $i \in \mathcal{P}$, is replaced by the model $\varepsilon(y_i - \theta_j) = 0$ or $\varepsilon(y_i - \theta_j x_i) = 0$, $i \in \mathcal{P}_j, j = 1, ..., k$. That is, now θ can vary from stratum to stratum. Generally, in practice, one cannot be sure if for the survey population at hand, the parameter θ has a "common" value for all individuals $i \in \mathcal{P}$. Theoretical as well as numerical investigations clearly indicate that the performance of the confidence intervals computed on the assumption of a common value of θ (*e.g.*, the ones based on the pivots $[g / \{ \hat{V}(g) \}^{1/2}]$ of section 4) is very susceptible even to "small deviations" of θ, from stratum to stratum. Of course it typically happens that in stratifying a population, a prior assessment of the mean values θ for different individuals leads to construction of strata $\mathcal{P}_j$ with differing mean values $\theta_j, j = 1, ..., k$.

For the reasons given above we propose the general use of confidence intervals based on the pivot (7.9), when there is no covariate; and confidence intervals based on the pivot

(7.11) for a covariate case. In the following illustrations the above confidence intervals are compared with conventional confidence intervals: These are obtained, in the case of no covariate, from the approximate $N(0, 1)$ pivot

$$\frac{\left\{\sum_{j=1}^{k} \frac{N_j}{N} \frac{1}{n_j} \sum_{i \in s_j} (y_i - \bar{Y})\right\}}{\left\{\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left(\frac{1}{n_j} - \frac{1}{N_j}\right) \frac{1}{(n_j-1)} \sum_{i \in s_j} (y_i - \bar{y}_j)^2\right\}^{1/2}}; \qquad (8.1)$$

in case of a covariate the approximate $N(0, 1)$ pivot is

$$\frac{\sum_{j=1}^{k} \frac{N_j}{N} \cdot \frac{1}{n_j} \sum_{i \in s_j} \left(y_i - \frac{\bar{Y}}{\bar{X}} x_i\right)}{\left[\sum_{j=1}^{k} \frac{N_j^2}{N^2} \cdot \left(\frac{1}{n_j} - \frac{1}{N_j}\right) \frac{1}{n_j - 1} \sum_{i \in s_j} \left\{(y_i - \bar{y}_j) - \frac{\bar{y}}{\bar{x}}(x_i - \bar{x}_j)\right\}^2\right]^{1/2}}. \qquad (8.2)$$

(Cochran 1977). In general we will refer to (7.9) and (7.11) as the new pivots and (8.1) and (8.2) as the conventional pivots.

Extensive simulation experiments were conducted to compare confidence intervals based on the new and the conventional pivots. However the results reported below are primarily for small samples. Here we have sixteen survey populations, each of which is divided into four strata; samples of sizes 2, 3, 4, 2 are drawn from the respective strata. Such samples of (total) sizes as small as 11, can bring out best, as in Tables 1 and 2 to follow, the superior performance of the new pivots over the conventional ones. To a lesser degree than for the small size samples just mentioned, the superiority of the new pivots over the conventional ones continues to hold for moderate size (25) samples, as in Table 3 and 4. Our unreported simulation studies included populations divided into 16 strata each, with total sample size of about 50. Even for such large samples the new pivots appear to perform better than the conventional ones. Eventually, of course, for very large sample sizes, the distinctions in performance between the two pivots, the new and the conventional, tend to disappear.

The sixteen survey populations (1) – (16) in Tables 1 and 2 below, except for populations (7), (8), are of sizes 1,000 each; populations (7), (8) are of sizes 2,000 each. Each one of the sixteen populations, as said before, is divided in 4 strata. Tables 1 and 2 each have six columns (i), (ii), ... (vi). Column (i) gives the population number (·). Column (ii) provides, corresponding to the four strata of the population (·), the superpopulation distributions from which the strata have been drawn. The distribution can be Chi-square (*C*), Normal (*N*), or Uniform (*U*). When there is no covariate as in Table 1, column (ii) refers to just the distribution of the variate *y*; on the other hand in Table 2, it refers to the distributions of both the variate *y* and the covariate *x*, with *y* (conditional on *x*) having mean θx. Column (iii) gives the sample sizes from different strata. Column (iv) shows the nominal coverage probability. Columns (v) and (vi) provide the actual coverage probabilities attained and the average length of the confidence intervals, under 4,000 simulations. Thus a typical horizontal line in Table 1, starting with (6) say, is to be read as follows. The four strata of the population (6) are drawn from the superpopulation distributions Normal, Chi-square, Normal, Chi-square respectively; the sample sizes from different strata are (2, 3, 4, 2) respectively. The interpretation of the columns (iv), (v), (vi) is straightforward. Unlike the populations (1) – (16) above, the populations (17) and (18) in Tables 3 and 4 are divided into 8 strata each, the population (17) being without a covariate and (18) with a covariate.

Table 1

(i)	(ii)	(iii)	(iv)	(v)		(vi)	
Population	Superpopulation distribution *y*	Sample sizes	Nominal coverage probability	Actual coverage probability		Average length	
				pivot (7.9)	pivot (8.1)	pivot (7.9)	pivot (8.1)
(1)	{N,C,U,C}	(2,3,4,2)	.95	.967	.86	19.83	11.33
(2)	{N,C,U,C}	(2,3,4,2)	.90	.90	.80	13.11	9.51
(3)	{N,N,N,N}	(2,3,4,2)	.90	.90	.807	4.85	3.52
(4)	{U,U,U,U}	(2,3,4,2)	.90	.90	.817	4.90	3.55
(5)	{N,C,N,C}	(2,3,4,2)	.95	.946	.82	34.34	19.62
(6)	{N,C,N,C}	(2,3,4,2)	.90	.866	.76	22.71	16.46
(7)	{N,U,C,N}	(2,3,4,2)	.95	.97	.869	20.76	11.80
(8)	{N,U,C,N}	(2,3,4,2)	.90	.908	.81	13.69	9.96

In the populations numbered (1) - (4) below the mean value θ is held fixed from stratum to stratum, $\theta = 100$, the standard deviation varies between 2.0 and $\sqrt{200.00}$. For the remaining populations, (5) to (8), the mean value θ varies from stratum to stratum, between $\theta = 100$ and $\theta = 400$

Table 2

(i) Population	(ii) Superpopulation distribution x	 y	(iii) Sample Sizes	(iv) Nominal coverage probability	(v) Actual coverage probability pivot (7.11)	 pivot (8.2)	(vi) Average length pivot (7.11)	 pivot (8.2)
(9)	{U,U,U,U}	{C,C,C,C}	(2,3,4,2)	.90	.879	.82	91.12	54.12
(10)	{U,U,U,U}	{C,C,C,C}	(2,3,4,2)	.95	.926	.876	113.49	64.49
(11)	{U,U,U,U}	{N,N,N,N}	(2,3,4,2)	.90	.88	.84	12.21	6.85
(12)	{C,C,C,C}	{N,N,N,N}	(2,3,4,2)	.90	.83	.83	7.07	6.84
(13)	{C,C,C,C}	{C,C,C,C}	(2,3,4,2)	.95	.926	.87	113.53	100.10
(14)	{C,C,C,C}	{C,C,C,C}	(2,3,4,2)	.90	.869	.84	35.89	33.01
(15)	{C,C,C,C}	{C,C,C,C}	(2,3,4,2)	.95	.92	.89	42.88	39.34
(16)	{C,C,C,C}	{U,C,C,U}	(2,3,4,2)	.95	.959	.909	31.17	26.68

In the Populations numbered (9) - (12) below the regression coefficient θ is held fixed for all strata, $\theta = 3$. For the remaining populations (13) - (15), the regression coefficients θ varies from stratum to stratum, between $\theta = 2$ and $\theta = 4$

Table 3

(i) Population	(ii) Superpopulation distribution y	(iii) Sample sizes	(iv) Nominal coverage	(v) Actual coverage probability pivot (7.9)	 pivot (8.1)	(vi) Average length pivot (7.9)	 pivot (8.1)
(17)	$\{N,U,C,U,C,N,U,U\}$	(2,3,4,2,3,4,3,4)	.95	.93	.889	12.76	10.94

In the population numbered (17) below, the mean value θ varies from stratum to stratum between $\theta = 100$ and $\theta = 800$

Table 4

(i) Population	(ii) Superpopulation distribution	(iii) Sample sizes	(iv) Nominal coverage	(v) Actual coverage probability pivot (7.11)	 pivot (8.2)	(vi) Average length pivot (7.11)	 pivot (8.2)
(18)	x: $\{C,C,C,C,C,C,C,C\}$ y: $\{N,U,N,C,C,C,C,N\}$	(2,3,4,2,3,4,3,4)	.95	.937	.90	22.80	20.89

In the population numbered (18) below, the regression coefficient θ varies from stratum to stratum between $\theta = 3$ and $\theta = 6$

9. CONCLUSIONS

The following conclusions are based on the theoretical investigations of the preceding sections and the simulation results reported in section 8, as well as many other simulation results, as mentioned earlier, not reported in this paper.

The situation when there is no covariate seems to be fairly clear from Tables 1 and 3 of section 8. For small samples the conventional confidence intervals, that is the ones based on the pivot (8.1), can be very misleading: The "asserted" probability of coverage can be very different than the "actual" one. Further, this gap between asserted and actual coverage probabilities for the conventional confidence intervals seems to increase as the variation in the stratum means increases. Interestingly, as noted in section 7, this increased variation in the stratum means can often be a result of stratifying a population into (internally) homogeneous strata for efficient point estimation. The confidence intervals based on the new pivot (7.9), as it can be seen from the Tables 1 and 3 of section 8, perform much better than the ones based on the conventional pivot (8.1). From our simulations, based on three distributions, namely Normal, Chi-square and Uniform, it seems that the comparison between performance of the new pivot (7.9) and the conventional one (8.1) depends on the distributions mostly through their variations of the mean values from stratum to stratum. Particularly, the comparison is not much affected by the variances or the forms of the distributions. This is to be expected from our underlying semi-parametric model, $\varepsilon(y_i - \theta_j) = 0, i \in \mathcal{P}_j, j = 1, \ldots, k$. This thus extends the conclusion previously drawn in the beginning of section 8. We emphasize here that the optimality of the estimating function g in (4.3) continues to hold even when θ varies from stratum to stratum.

For large samples, according to our simulation results mentioned earlier (unreported here) the difference between the two sets of confidence intervals, one based on the pivot (7.9) and the other on (8.1), tend to diminish. This also is in line with the theory.

Tables 2 and 4 of section 8 provide results concerning confidence intervals for populations admitting a covariate. Here a comparison of the performances of the new pivot (7.11) and the conventional one (8.2) is rather subtle. We consider two situations: One, when the regression coefficient θ, is the same for all strata; Two, when θ varies (though not very much) from stratum to stratum. Only in the former situation is the estimating function g in (4.11) optimal. For this situation (*i.e.*, same θ for all strata), which is mostly of academic interest, the pivot given at the end of section 4, a sour simulation studies (unreported here) show, performs very well. Situation two, above, is more realistic. Hence it is practically very important to study the performance of the estimating function g, that is the performance of the confidence intervals based on the new pivot (7.11), when θ varies from stratum to stratum. Under this situation, it is clear from Tables 2 and 4 that the confidence intervals based on the new pivot (7.11) provide "actual" coverage probabilities closer to the "asserted" ones than the confidence intervals based on the conventional pivot (8.2). Also under situation one, *i.e.*, the same θ for all strata, as Table 2 indicates, the performance of the new pivot (7.11) is at least as good as the conventional pivot (8.2). The phenomenon seems to be more striking as more variation in the covariate values is introduced. Actually, as the covariate values within each stratum tend to be uniform the difference between the performances of the two pivots, the new (7.11) and the conventional (8.2), tends to diminish. The difference also diminishes as the sample sizes go on increasing.

A referee has suggested that, since for a single stratum case our new pivot (7.11) closely resembles Fieller's pivot, referred to by Cochran (1977), we comment on Cochran's observations: for some parametric distributions (for example bivariate normality of (x, y) with one of the means close to zero), the confidence intervals for the ratio of means based on Fieller's pivot may have undesirable properties (concerning probability of coverage and interval length) in comparison to the confidence intervals based on the conventional pivot (8.2). In the survey sampling context such circumstances would be exceptional, as indicated by our simulation results. Moreover, our new pivot (7.11) has "validity" for a semi-parametric model; underlying this is a large class of parametric models.

ACKNOWLEDGEMENTS

We are grateful to A.C. Singh for valuable comments on an earlier draft of this paper. Our thanks are due to Jiahua Chen and Lianxiang Wang for their advice and assistance in computations.

APPENDIX

The variance of the estimating function g in (5.3) namely $V(g)$, is $E(g^2)$ since $E(g) = 0$. Further

$$E(g^2) = E\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{\sum_{c\in s_j}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\Big/ n_j\right\}^2 +$$

$$\sum_{\substack{j\neq j'\\ j,j'=1}} \frac{N_j N_{j'}}{N^2} E\left[\left\{\sum_{c\in s_j}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\Big/ n_j\right\}\left\{\sum_{c\in s_{j'}}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\Big/ n_{j'}\right\}\right],$$

$$= A + B \text{ say.} \tag{I}$$

Then

$$A = \sum_{j=1}^{k} \frac{N_j^2}{N^2} E\left\{\sum_{c\in s_j}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)^2\Big/ n_j^2\right\} +$$

$$\sum_{j=1}^{k} \frac{N_j^2}{N^2} E\left\{\sum_{\substack{c\neq c'\\ c,c'\in s_j}} E_{c,c'}\left(\hat{Y}_c - \frac{Y}{X}X_c\right)\left(\hat{Y}_{c'} - \frac{Y}{X}X_{c'}\right)\Big/ n_j^2\right\}, \tag{II}$$

where $E_{c,c'}$ denotes the expectation holding the clusters c,c' fixed. Now as stated in the beginning of section 5, the sampling designs for different clusters are independent. Further, the second term in the r.h.s. of "A" is equal to

$$\sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{E\left[\frac{1}{n_j}\sum_{c\in s_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right]\right.$$

$$\left. - E\left[\frac{1}{n_j^2}\sum_{c\in s_j}\left(Y_c - \frac{Y}{X}X_c\right)^2\right]\right\}$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{\left(\frac{1}{n_j} - \frac{1}{N_j}\right)\frac{1}{N_j - 1}\sum_{c\in\mathcal{P}_j}\left[(Y_c - \bar{Y}_j) - \frac{Y}{X}(X_c - \bar{X}_j)\right]^2\right.$$

$$\left. + (\bar{Y}_j - \frac{Y}{X}\bar{X}_j)^2 - \frac{1}{n_j N_j}\sum_{c\in\mathcal{P}_j}(Y_c - \frac{Y}{X}X_c)^2\right\}$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{[(\frac{1}{n_j} - \frac{1}{N_j})\frac{1}{N_j - 1}\sum_{c\in\mathcal{P}_j}(Y_c - \frac{Y}{X}X_c)^2]\right.$$

$$\left. + \frac{N_j(n_j - 1)}{n_j(N_j - 1)}(\bar{Y}_j - \frac{Y}{X}\bar{X}_j)^2 - \frac{1}{n_j N_j}\sum_{c\in\mathcal{P}_j}(Y_c - \frac{Y}{X}X_c)^2\right\}$$

$$= \sum_{j=1}^{k} \frac{N_j^2}{N^2}\left\{\frac{N_j(n_j - 1)}{n_j(N_j - 1)}(\bar{Y}_j - \frac{Y}{X}\bar{X}_j)^2\right.$$

$$\left. - \frac{n_j - 1}{N_j(N_j - 1)}\frac{1}{n_j}\cdot\sum_{c\in\mathcal{P}_j}(Y_c - \frac{Y}{X}X_c)^2\right\} \tag{III}$$

where $\bar{Y}_j = \sum_{c \in \mathcal{P}_j} Y_c / N_j$ and $\bar{X}_j = \sum_{c \in \mathcal{P}_j} X_c / N_j$, $j = 1, \ldots, k$. The term B in $E(g^2)$ simplifies as

$$B = \left[\sum_{j=1}^{k} \frac{N_j}{N} \left(\bar{Y}_j - \frac{Y}{X} \bar{X}_j \right) \right]^2 - \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\bar{Y}_j - \frac{Y}{X} \bar{X}_j \right)^2$$

$$= - \sum_{j=1}^{k} \frac{N_j^2}{N^2} \left(\bar{Y}_j - \frac{Y}{X} \bar{X}_j \right)^2 . \qquad \text{(IV)}$$

Note that because of the superpopulation model $\varepsilon(y_i - \theta x_i) = 0$, $i \in \mathcal{P}$ in (III) and (IV) the term

$$\left(\bar{Y}_j - \frac{Y}{X} \bar{X}_j \right)^2 \text{ is } O\left(\frac{1}{N_j} \right), \; j = 1, \ldots, k.$$

It can thus be shown that second term of A and the term B are both of order $O(1/N)$, while the first term of A is of order $O(N^2/n^3)$. Hence from (I) – (IV) for large strata sizes N_j, $j = 1, \ldots, k$ we have approximations to the variance $V(g)$ and its estimate $\hat{V}(g)$ as in (5.4) and (5.5). These approximations are valid, regardless of the superpopulation model just mentioned, provided the sample sizes n_j in addition to the strata sizes N_j, $j = 1, \ldots, k$ are sufficiently large.

REFERENCES

BASU, D. (1958). On sampling with and without replacement. *Sankhyā*, 20, 287-294.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 39, 1035-1043.

CHAUDHURI, A., and VOS, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. Amsterdam: North Holland.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.

DEMING, W.E. (1950). *Some Theory of Sampling*. New York: John Wiley.

EFRON, B., and HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected fisher information (with discussion.) *Biometrika*, 65, 457-487.

FISHER, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, 700-725.

GODAMBE, V.P. (1976). A historical perspective of recent developments in the theory of sampling from actual populations. *Journal of the Indian Society of Agricultural Statistics*, 28, 1-12.

GODAMBE, V.P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika*, 72, 419-428.

GODAMBE, V.P. (1991). Orthogonality of Estimating functions and nuisance parameters. *Biometrika*, 78, 143-151.

GODAMBE, V.P. (1995). Estimation of parameters in survey sampling: Optimality. *Canadian Journal of Statistics*, 23, 227-243.

GODAMBE, V.P. (1997). Estimation of parameters in survey sampling. *1996 Proceedings of the Survey Methods Section, Statistical Society of Canada.*

GODAMBE, V.P., and HEYDE, C.C. (1987). Quasi-likelihood and optimal estimation. *International Statistical Review*, 55, 231-244.

GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationship and estimation. *International Statistical Review*, 54, 127-138.

GODAMBE, V.P., and THOMPSON, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). *Journal of Statistical Planning and Inference*, 22, 137-172.

HÁJEK, J. (1959). Optimum strategy and other problems in probability sampling. *Casopis Pro Péstování Matematiky*, 84, 387-423.

MACH, L. (1988). The Use of Estimating Functions for Confidence Interval Construction: The Case of the Population Mean. Working Paper No. BSMD-88-028 E Methodology Branch, Statistics Canada.

NEYMAN, J. (1934). On two different aspects of representative method: the method of stratified sampling and the method of purposive selection, (with discussion). *Journal of the Royal Statistical Society*, 97, 558-652.

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of Royal Society*. Series A, 236, 333-380.

ROYALL, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review* 54, 221-226.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SMITH, T.M.F. (1997). Social surveys and social science, (with discussion). *Canadian Journal of Statistics*, 25, 23-44.

VINOD, H.D. (1998). Foundations of statistical inference based on numerical roots of robust pivot functions. *Journal of Econometrics*, 81, 387-396.

WILKS, S.S. (1938). Shortest average confidence intervals from large samples. *Annals of Mathematical Statistics*, 9, 166-175.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

YUNG, W., and RAO, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 175-186
Statistics Canada

Some Recent Advances in Model-Based Small Area Estimation

J.N.K. RAO[1]

ABSTRACT

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area estimators. Traditional area-specific direct estimators do not provide adequate precision because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow strength from related areas; in particular, model-based indirect estimators. Ghosh and Rao (1994) provided a comprehensive review and appraisal of methods for small area estimation, covering the literature to 1992-1993. This paper supplements Ghosh and Rao (1994) by covering the literature over the past five years or so on model-based estimation. In particular, we cover several small area models and empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB) and hierarchical Bayes (HB) methods applied to these models. We also present several recent applications of small area estimation.

KEY WORDS: Empirical Bayes; Hierarchical Bayes; Small Area Models.

1. INTRODUCTION

Sample surveys are used to provide estimates not only for the total population but also for a variety of subpopulations (domains). "Direct" estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains. But sample sizes in small domains, particularly small geographical areas, are rarely large enough to provide direct estimates for specific small domains. For example, the U.S. Third National Health and Nutrition Examination Survey was designed to provide direct estimates with acceptable precision for domains classified by race, ethnicity and age. But, to have a large enough sample to support reliable direct estimates for, say, all states is seldom possible, and for all subareas like counties is almost never possible. In this example, states/counties may be regarded as "small areas" because the area-specific sample sizes are small (or even zero). In making estimates for such small areas it is necessary to "borrow strength" from related areas to form "indirect" estimators that increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as recent census counts and current administrative records. Indirect estimators based on implicit models include synthetic and composite estimators, while those based on explicit models incorporating area-specific effects include empirical Bayes (EB), empirical best linear unbiased prediction (EBLUP) and hierarchical Bayes (HB) estimators.

Ghosh and Rao (1994) presented a comprehensive overview and appraisal of methods for small area estimation, covering the literature to 1992-1993. We refer the reader to Schaible (1996) for an excellent account of the use of indirect estimators in U.S. Federal Programs.

Ghosh and Rao (1994) provided a list of symposia and workshops on small area estimation that have been organized in recent years. We update that list by the following: (i) Conference on Small Area Estimation, U.S. Bureau of the Census, Washington, D.C., March 26-27, 1998; and (ii) International Satellite Conference on Small Area Estimation, Riga, Latvia, August 20-21, 1999. Short courses have also been organized: (i) "Small Area Estimation" by J.N.K. Rao, W.A. Fuller, G. Kalton and W.L. Schaible, organized by the Joint Program in Survey Methodology and the Washington Statistical Society, Washington, D.C., May 22-23, 1995; and (ii) "Introduction to Small Area Estimation" by J.N.K. Rao, organized by the International Association of Survey Statisticians, Riga, Latvia, August 19, 1999. In addition, numerous invited and contributed sessions on small area estimation have been organized at recent professional statistical meetings, including the American Statistical Association Annual Meetings and the International Statistical Institute bi-annual sessions.

Singh, Gambino and Mantel (1994) discussed survey design issues that have an impact on small area statistics. In particular, they presented an excellent illustration of compromise sample size allocations to satisfy reliability requirements at the provincial level as well as sub provincial level. For the Canadian Labour Force Survey with a monthly sample of 59,000 households, optimizing at the provincial level yields a coefficient of variation (CV) for "unemployed" as high as 17.7% for some Unemployment Insurance (UI) regions. On the other hand, a two-step allocation with 42,000 households allocated at the first step to get reliable provincial estimates and the remaining 17,000 households allocated in the second step to produce best possible UI region estimates reduces the worst case of 17.7% CV for UI regions to 9.4% at the expense of a small increase in CV at the provincial and national levels: CV for Ontario increases from 2.8% to 3.4%

[1] J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

and for Canada from 1.36% to 1.51%. Preventive measures, such as compromise sample allocations, should be taken at the design stage, whenever possible, to ensure precision for domains like the UI region. But even after taking such measures sample sizes may not be large enough for direct estimates to provide adequate precision for all small areas of interest. As noted before, sometimes the survey is deliberately designed to oversample specific areas (domains) at the expense of small samples or even no samples in other areas of interest.

This paper supplements Ghosh and Rao (1994) by covering the literature over the past five years or so on model-based small area estimation; in particular, on empirical best linear unbiased prediction (EBLUP), empirical Bayes (EB and hierarchical Bayes (HB) methods and their applications.

2. SMALL AREA MODELS

It is now generally accepted that when indirect estimates are to be used they should be based on explicit models that relate the small areas of interest through supplementary data such as last census data and current administrative data. An advantage of the model approach is that it permits validation of models from the sample data. Interesting work on traditional indirect estimates (synthetic, sample-size dependent *etc.*), however, is also reported in the recent literature (see *e.g.*, Falorsi, Falorsi and Russo 1994; Chaudhuri and Adhikary 1995; Schaible 1996; Marker 1999).

Small area models may be broadly classified into two types: area level and unit level.

2.1 Area Level Models

Area-specific auxiliary data, $\mathbf{x}_i$, are assumed to be available for the sampled areas i (= 1, ..., m) as well as the nonsampled areas. A basic area level model assumes that the population small area mean $\bar{Y}_i$ or some suitable function $\theta_i = g(\bar{Y}_i)$, such as $\theta_i = \log(\bar{Y}_i)$, is related to $\mathbf{x}_i$ through a linear model with random area effects v_i:

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i, \qquad i = 1, ..., m \tag{2.1}$$

where β is the p-vector of regression parameters and the v_i's are uncorrelated with mean zero and variance σ_v^2. Normality of the v_i is also often assumed. The model (2.1) also holds for the non sampled areas. It is also possible to partition the areas into groups and assume separate models of the form (2.1) across groups.

We assume that direct estimators $\hat{\bar{Y}}_i$ of $\bar{Y}_i$ are available whenever the area sample size $n_i \geq 1$. It is also customary to assume that

$$\hat{\theta}_i = \theta_i + e_i \tag{2.2}$$

where $\hat{\theta}_i = g(\hat{\bar{Y}}_i)$ and the sampling errors e_i are independent $N(0, \psi_i)$ with known ψ_i. Combining this sampling model with the "linking" model (2.1), we get the well-known area level linear mixed model of Fay and Herriot (1979):

$$\hat{\theta}_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i + e_i. \tag{2.3}$$

Note that (2.3) involves both design-based random variables e_i and model-based random variables v_i. In practice, sampling variances ψ_i are seldom known, but smoothing of estimated variances $\hat{\psi}_i$ is often done to get stable estimates ψ_i^* which are then treated as the true ψ_i. Other methods of handling unknown ψ_i are mentioned in section 3.4. An advantage of the area-level model (2.3) is that the survey weights are accounted for through the direct estimators $\hat{\theta}_i$.

The assumption $E(e_i \mid \theta_i) = 0$ in the sampling model (2.2) may not be valid if the sample size n_i is small and θ_i is a nonlinear function of the total Y_i, even if the direct estimator $\hat{Y}_i$ is design-unbiased, *i.e.*, $E(\hat{Y}_i \mid Y_i) = Y_i$. A more realistic sampling model is given by

$$\hat{Y}_i = Y_i + e_i^* \tag{2.4}$$

with $E(e_i^* \mid Y_i) = 0$, *i.e.*, $\hat{Y}_i$ is design-unbiased for the total Y_i. In this case, however, we cannot combine (2.4) with the linking model to produce a linear mixed model. As a result, standard results in linear model theory do not apply, unlike in the case of (2.3). Alternative methods to handle this case are needed (see section 4.1).

The basic area level model has been extended to handle correlated sampling errors, spatial dependence of random small area effects, vectors of parameters θ_i (multivariate case), time series and cross-sectional data and others (see Ghosh and Rao 1994). We discuss some of the recent models for combining cross-sectional and time series data. Suppose θ_{it} denotes a parameter of interest for small area i at time t and $\hat{\theta}_{it}$ is a direct estimator of θ_{it}. Ghosh, Nangia and Kim (1996) assumed the sampling model $\hat{\theta}_{it} \mid \theta_{it} \overset{\text{ind}}{\sim} N(\theta_{it}, \psi_{it})$ with known sampling variances ψ_{it}, and the linking model

$$\theta_{it} \mid \mathbf{u}_t \sim N(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{u}_t, \sigma_t^2) \tag{2.5}$$

and

$$\mathbf{u}_t \mid \mathbf{u}_{t-1} \sim N(\mathbf{u}_{t-1}, \mathbf{W}) \tag{2.6}$$

with known auxiliary variables $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$; they have actually studied the multivariate case θ_{it}. Note that (2.6) is the well-known random walk model. The above model has the following limitations: (i) Independence of $\hat{\theta}_{it}$'s over t for each i may not be realistic because estimates are typically correlated over time. (ii) The linking model (2.5) does not include area-specific random effects. As a result, it is likely to produce oversmooth estimates. Rao and Yu (1992, 1994) proposed more realistic sampling and linking models. They assumed the sampling model

$$\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{\theta}_i \overset{\text{ind}}{\sim} N(\boldsymbol{\theta}_i, \boldsymbol{\psi}_i) \tag{2.7}$$

with known sampling covariance matrix ψ_i, and the linking model

$$\theta_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} + v_i + u_{it} \qquad (2.8)$$

with $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$ and independent of u_{it}'s which are assumed to follow an AR(1) model:

$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, \quad |\rho| < 1 \qquad (2.9)$$

with $\varepsilon_{it} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, where $\hat{\theta}_i = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{iT})'$ and $\theta_i = (\theta_1, \ldots, \theta_{iT})'$. Models of the form (2.7)-(2.9) have been extensively studied in the econometric literature, ignoring sampling errors, *i.e.*, treating $\hat{\theta}_{it}$ as θ_{it}. The above sampling model permits correlations among sampling errors over time and the linking model (1.9) includes both area-specific random effects v_i and area by time specific random effects u_{it}. Datta, Lahiri and Lu (1994), following Rao and Yu (1992), used the same sampling model (2.7) but assumed the following linking model:

$$\theta_{it} \mid v_i, \mathbf{u}_t \sim N(\mathbf{x}_{it}^T \boldsymbol{\beta}_i + v_i + \mathbf{z}_{it}^T \mathbf{u}_t, \sigma_i^2) \qquad (2.10)$$

where $\boldsymbol{\beta}_i$'s and σ_i^2's are random and $\mathbf{u}_t$ follows the random walk model (2.6). This model allows area-specific random effects v_i and random slopes $\boldsymbol{\beta}_i$, but does not contain area by time specific random effects u_{it}. Datta, Lahiri and Maiti (1999) used the Rao-Yu sampling and linking models (2.7) and (2.8) but replaced the AR(1) model (2.9) by a random walk model given by (2.9) with $\rho = 1$. Datta, Lahiri, Maiti and Lu (1999) considered a similar model but added extra terms to $\mathbf{x}_{it}^T \beta + v_i$ to reflect seasonal variation in their application to estimating unemployment rates for the U.S states. Singh, Mantel and Thomas (1994) also used time series/cross-sectional models, but assumed that the sample errors are uncorrelated over time.

Area level models have also been used in the context of disease mapping or estimating regional mortality and disease rates, as noted by Ghosh and Rao (1994). A simple model assumes that the observed small area disease counts $y_i \mid \theta_i \overset{\text{ind}}{\sim}$ Poisson $P(n_i \theta_i)$ and $\theta_i \overset{\text{i.i.d.}}{\sim}$ gamma $G(a, b)$, where θ_i is the true incidence rate and n_i is the number exposed in area i. Maiti (1998) used $\beta_i = \log \theta_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ instead of $\theta_i \overset{\text{i.i.d.}}{\sim} G(a, b)$. He also considered a spatial dependence model for β_i's, using conditional autoregression (CAR) that relates each β_i to a set of neighbourhood areas of area i; see also Ghosh, Natarajan, Kim and Walker (1997). Lahiri and Maiti (1996) modelled age-group specific area disease counts y_{ij}, using Clayton and Kaldor's (1987) approach. They assumed that $y_{i\cdot} = \sum_j y_{ij} \mid \theta_i \overset{\text{ind}}{\sim} P(e_i \theta_i)$ and $\theta_i \overset{\text{i.i.d.}}{\sim} G(a, b)$, where $e_i = \sum_j \psi_j n_{ij}$ is the expected number of deaths in area i, ψ_j is the j-th group effect assumed to be known and n_{ij} is the number exposed in the j-th age group and area i. Nandram, Sedransk and Rickle (1998) assumed that $y_{ij} \mid \theta_{ij} \overset{\text{ind}}{\sim} P(n_{ij}\, \theta_{ij})$ and $\log \theta_{ij} = \mathbf{x}_j' \beta + v_i$ with $v_i \overset{\text{ind}}{\sim} N(0, \sigma^2)$, where θ_{ij} is the area/age-specific mortality rate and $\mathbf{x}_j$ is a vector of covariates for age group j. They also considered random slopes $\boldsymbol{\beta}_i$ in the linking model.

2.2 Unit Level Models

A basis unit level population model assumes that the unit y-values y_{ij}, associated with the units j in the areas i, are related to auxiliary variables $\mathbf{x}_{ij}$ through a one-way nested error regression model

$$y_{ij} = \mathbf{x}_{ij}' \boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, N_i;\ i = 1, \ldots, m \qquad (2.11)$$

where $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$ are independent of $e_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2)$ and N_i is the number of population units in the i-th area. The parameters of interest are the totals Y_i or the means $\bar{Y}_i$.

The model (2.11) is appropriate for continuous variables y. To handle count or categorical (*e.g.*, binary) y-variables, generalized linear mixed models with random small area effects, v_i, are often used. Ghosh, Natarajan, Stroud and Carlin (1998) assumed models of the form: (i) Given θ_{ij}'s, the y_{ij}'s are independent and belong to the exponential family with canonical parameter θ_{ij}; (ii) Linking model $g(\theta_{ij}) = \mathbf{x}_{ij}' \beta + v_i$ where $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$ and $g(\cdot)$ is a strictly increasing function. The linear mixed model (2.11) is a special case of this class with $g(a) = a$. The logistic function $g(a) = \log\,[a/(1-a)]$ is often used for binary y (see *e.g.*, Farrell, McGibbon and Tomberlin 1997) although probit functions can also be used and offer certain advantages for hierarchical Bayes (HB) inference (Das, Rao and You 1999).

The sample data $\{y_{ij}, \mathbf{x}_{ij}, j = 1, \ldots, n_i;\ i = 1, \ldots, m\}$ is assumed to obey the population model. This implies that the sample design is ignorable or selection bias is absent which is satisfied by any equal probability sampling method within areas. For more general designs, the sample indicator variable, a_{ij}, should be unrelated to y_{ij}, conditional on x_{ij}. Model-based estimators for unit level models do not depend on the survey weights, $\tilde{w}_{ij}$, so that design-consistency as n_i increases is forsaken except when the design is self-weighting, *i.e.*, $\tilde{w}_{ij} = \tilde{w}$, as in the case of equal probability sampling. The area level model (2.3) is free of these limitations but assumes that the sample variances ψ_i are known; if ψ_i's are assumed unknown the model becomes nonidentifiable or nearly nonidentifiable leading to highly unstable estimates of the parameters. The unit level model is free of the latter difficulty and survey weights can also be incorporated using model-assisted estimators; see the paragraph containing equation (3.8).

Various extensions of the basic unit level models have been studied over the past five years or so. Stukel and Rao (1999) studied two-way nested error regression models which are appropriate for two-stage sampling within small areas. Following Kleffe and Rao (1992), Arora and Lahiri (1997) studied unit level models of the form (2.11) with random error variances σ_i^2 such that $\sigma_i^{-2} \overset{\text{i.i.d.}}{\sim} G(a, b)$; Kleffe and Rao (1992) assumed the existence of only mean and variance of σ_i^2, without specifying a parametric distribution on σ_i^2. Datta, Day and Basawa (1999) extended the unit level model (2.11) to the multivariate case y_{ij}, following Fuller and Harter (1987). This extension leads to a multivariate nested error regression model. Moura and Holt (1999) generalized (2.11) to allow some or all of the regression coefficients to be random and to depend on area level auxiliary variables, thus effectively integrating the use of unit level and area level covariates into a single model. You and Rao (1999a) also studied similar two-level models.

Malec, Davis and Cao (1996, 1999) and Malec, Sedransk, Moriarity and LeClere (1997) studied the binary case, using logistic linear mixed models with random slopes to link the small areas. Raghunathan (1993) specified only the first two moments of y_{ij}'s conditional on small area means θ_i's and the first moment of θ_i as $\tau_i = h(\mathbf{z}_i' \beta)$ for known inverse "link" function $h(\cdot)$ and the second moment of θ_i is allowed to depend on τ_i.

Many of the small area linear mixed models studied in the literature are special cases of the following general linear mixed model with a block diagonal covariance structure, sometimes called longitudinal mixed linear models (Prasad and Rao 1990; Datta and Lahiri 1997):

$$\mathbf{y}_i^* = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{v}_i + \mathbf{e}_i, \quad i = 1, ..., m \tag{2.12}$$

where $\mathbf{v}_i \overset{\text{ind}}{\sim} (\mathbf{0}, \mathbf{G}_i(\tau))$ and independent of $\mathbf{e}_i \overset{\text{ind}}{\sim} (\mathbf{0}, \mathbf{R}_i(\tau))$. For example, the basic area level (2.3) is of the form (2.12) with $\mathbf{y}_i^* = \hat\theta_i, \mathbf{Z}_i = \mathbf{1}, \mathbf{G}_i(\tau) = \sigma_v^2$ and $\mathbf{R}_i(\tau) = \psi_i$. Das, Rao and You (1999) studied general mixed ANOVA models of the form

$$\mathbf{y}^* = \mathbf{X} \boldsymbol{\beta} + \mathbf{Z}_1 \mathbf{v}_1 + \cdots + \mathbf{Z}_q \mathbf{v}_q + e_i \tag{2.13}$$

where $\mathbf{Z}_i$ consists of only 0's and 1's such that there is exactly one 1 in each row and at least one 1 in each column, $\mathbf{v}_i \overset{\text{ind}}{\sim} (\mathbf{0}, \sigma_i^2 \mathbf{I})$ and independent of $e \sim (\mathbf{0}, \sigma^2 \mathbf{I})$. This model relaxes the assumption of a block diagonal covariance structure.

Ghosh and Rao (1994) reviewed some work on model diagnostics for models involving random effects. Jiang, Lahiri and Wu (1998) developed a chi-squared test for checking the normality of the random effects v_i and the errors e_{ij} in the basic unit level sample model $y_{ij} = \mathbf{x}_{ij}' \beta + v_i + e_{ij}, j = 1, \ldots, n_i; i = 1, \ldots, m$.

3. MODEL-BASED INFERENCE: BASIC AREA-LEVEL MODEL

EBLUP, EB and HB methods have played a prominent role for model-based small area estimation. EBLUP is applicable for linear mixed models whereas EB and HB are more generally valid. EBLUP point estimators do not require distributional assumptions, but normality of random effects is often assumed for estimating the mean squared error (MSE) of the estimators. Also, EBLUP and EB estimators are identical under normality and nearly equal to the HB estimator, but measures of variability of the estimators may be different. To illustrate the methods, we focus on the basic area level model (2.3), which is extensively used in practice. Various extensions of the basic area-level and unit level models are studied in section 4.

3.1 EBLUP Method

Appealing to general results for linear mixed models, the BLUP estimator of θ_i under (2.3) is given by

$$\tilde\theta_i(\sigma_v^2) = \gamma_i \hat\theta_i + (1 - \gamma_i)\mathbf{x}_i' \tilde{\boldsymbol{\beta}}(\sigma_v^2) \tag{3.1}$$

where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$ and $\tilde{\boldsymbol{\beta}}(\sigma_v^2)$ is the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ with weights $(\sigma_v^2 + \psi_i)^{-1}$. It follows from (3.1) that the BLUP estimator is a weighted combination of the direct estimator $\hat\theta_i$ and the regression synthetic estimator $\mathbf{x}_i' \tilde{\boldsymbol{\beta}}(\sigma_v^2)$. The result (3.1) does not require the normality of v_i and e_i. Since σ_v^2 is unknown, we replace it by a suitable estimator $\hat\sigma_v^2$ to obtain a two-step or EBLUP estimator $\tilde\theta_i = \tilde\theta_i(\hat\sigma_v^2)$. The estimator of total Y_i is taken as $g^{-1}(\tilde\theta_i) = h(\tilde\theta_i)$. One could use either the method of fitting constants (not requiring normality) or the restricted maximum likelihood (REML) method under normality to estimate σ_v^2. Jiang (1996) showed that REML estimators of variance components in linear mixed models remain consistent under deviations from normality. Therefore, $\tilde\theta_i$ with REML estimator of σ_v^2 is also asymptotically valid under nonnormality.

As noted in section 2.1, EBLUP estimation is not applicable if the sampling model (2.2) is changed to the more realistic model (2.4).

A measure of variability associated with EBLUP estimator is given by its MSE, but no closed form for MSE exists except in some special cases. As a result, considerable attention has been given in recent years to obtain accurate approximations to the MSE of EBLUP estimators. An accurate approximation to $\text{MSE}(\tilde\theta_i) = E(\tilde\theta_i - \theta_i)^2$, for large m, under normality is given by

$$\text{MSE}(\tilde\theta_i) \approx g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2) + g_{3i}(\sigma_v^2) \tag{3.2}$$

where

$$g_{1i}(\sigma_v^2) = \gamma_i \psi_i, \tag{3.3}$$

$$g_{2i}(\sigma_v^2) = (1 - \gamma_i)^2 \mathbf{x}_i' \left[\sum_i \mathbf{x}_i \mathbf{x}_i' / (\sigma_v^2 + \psi_i) \right]^{-1} \mathbf{x}_i, \tag{3.4}$$

$$g_{3i}(\sigma_v^2) = \left[\psi_i^2 / (\sigma_v^2 + \psi_i)^4 \right] E(\hat\theta_i - \mathrm{x}_i' \boldsymbol{\beta})^2 \bar{V}(\hat\sigma_v^2), \tag{3.5}$$

$$= \left[\psi_i^2 / (\sigma_v^2 + \psi_i)^2 \right] \bar{V}(\hat\sigma_v^2) \tag{3.6}$$

and $\bar{V}(\hat\sigma_v^2)$ is the asymptotic variance of $\hat\sigma_v^2$ (Prasad and Rao 1990). The leading term $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ is of order $O(1)$

whereas $g_{2i}(\sigma_v^2)$, due to estimating β, and $g_{3i}(\sigma_v^2)$, due to estimating σ_v^2, are both of order $O(m^{-1})$, for large m. Note that the leading term shows that $\text{MSE}(\tilde{\theta}_i)$ can be substantially smaller than $\text{MSE}(\hat{\theta}_i)$ under the model (2.3) when γ_i is small or the model variance σ_v^2 is small relative to the sampling variance ψ_i. The success of small area estimation, therefore, largely depends on getting good auxiliary information $\{\mathbf{x}_i\}$ that leads to a small model variance relative of ψ_i. Of course, one should also make a thorough validation of the assumed model.

An estimator of $\text{MSE}(\tilde{\theta}_i)$, correct to the same order of approximation as (3.2), is given by

$$\text{mse}(\tilde{\theta}_i) \approx g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2), \tag{3.7}$$

i.e., the bias of (3.7) is of lower order than m^{-1} for large m. The approximation (3.7) is valid for both the method of fitting constants estimator and the REML estimator, but not for the ML estimator of σ_v^2 (Datta and Lahiri 1997; Prasad and Rao 1990). Using the fitting of constants estimator, Lahiri and Rao (1995) showed that (3.7) is robust to nonnormality of the small area effects v_i in the sense that approximate unbiasedness remains valid. Note that the normality of sampling errors e_i is still assumed but it is less restrictive due to the central limit theorem's effect on the direct estimators $\hat{\theta}_i$.

A criticism of the MSE estimator (3.7) is that it is not area-specific in the sense that it does not depend on $\hat{\theta}_i$ although $\mathbf{x}_i$ involved through (3.4). But it is easy to find other choices using the form (3.5) for $g_{3i}(\sigma_v^2)$. For example, we can use

$$\begin{aligned}\text{mse}_1(\hat{\theta}_i) &= g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + g_{3i}(\hat{\sigma}_v^2)\\ &\quad + [\psi_i^2/(\hat{\sigma}_v^2 + \psi_i)^4]\,(\hat{\theta}_i - \mathbf{x}_i'\hat{\beta})^2 h_i(\hat{\sigma}_v^2),\end{aligned} \tag{3.8}$$

where $\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2)$ and $h_i(\sigma_v^2) = \bar{V}(\hat{\sigma}_v^2) = 2m^{-2}\sum_i(\sigma_v^2 + \psi_i)^2$ for the fitting of constants estimator $\hat{\sigma}_v^2$ (Rao 1998). The last term of (3.8) is less stable than $g_{3i}(\hat{\sigma}_v^2)$ but it is of lower order than the leading term $g_{1i}(\hat{\sigma}_v^2)$.

Rivest and Belmonte (1999) obtained an unbiased estimator of the conditional MSE of the EBLUP estimator $\tilde{\theta}_i = \tilde{\theta}_i(\hat{\sigma}_v^2)$ for the basic area level model, assuming only the sampling model, *i.e.*, conditionally given θ_i's. Hwang and Rao (1987) obtained similar results and showed empirically that the model-based estimator of MSE, (3.7), is much more stable than the unbiased estimator and that it tracks the conditional MSE quite well even under moderate violations of the assumed linking model (2.1). Only in extreme cases, such as large outliers θ_i, the model-based estimator might perform poorly compared to the unbiased estimator.

3.2 EB Method

In the EB approach to the basic area level model, given by (2.1) and (2.2), the conditional distribution of θ_i given $\hat{\theta}_i$ and model parameters β and σ_v^2, denoted $f(\theta_i \mid \hat{\theta}_i, \beta, \sigma_v^2)$. is first obtained. The model parameters are estimated from the marginal distribution of $\hat{\theta}_i$'s, and inferences are then based on the estimated conditional (or posterior) distribution of θ_i, $f(\theta_i \mid \hat{\theta}_i, \hat{\beta}, \hat{\sigma}_v^2)$. In particular, the mean of the estimated posterior distribution is the EB estimator $\tilde{\theta}_i^{\text{EB}}$. Under normality, $\tilde{\theta}_i^{\text{EB}}$ is identical to the EBLUP estimator $\tilde{\theta}_i$, but the EB approach is applicable generally for any joint distribution. It should be noted that the EB approach is essentially frequentist because it uses only the sampling model and the linking model which can be validated from the data; no priors on the model parameters are involved unlike in the HB approach.

As a measure of variability of $\tilde{\theta}_i^{\text{EB}}$, the variance of the estimated posterior is used. Under normality, it is given by $g_{1i}(\hat{\sigma}_v^2) = \hat{\gamma}_i\psi_i$ which leads to severe underestimation of true variability as measured by MSE. Laird and Louis (1987) proposed a parametric bootstrap method to account for the variability in $\hat{\beta}$ and $\hat{\sigma}_v^2$, but Butar and Lahiri (1997) showed that it is not second-order correct, *i.e.*, its bias involves terms of order m^{-1}, unlike the bias of (3.7) or (3.8). By correcting this bias, they obtained an estimator which is identical to the area-specific MSE estimator (3.8). Therefore, corrected EB and EBLUP lead to the same result under normality.

3.3 HB Method

The HB approach has been extensively used for small area estimation. It is straightforward, inferences are "exact" and it can handle complex problems using recently developed Monte Carlo Markov Chain (MCMC) methods, such as the Gibbs sampler. A prior distribution on the model parameters (also called hyper parameters) is specified and the posterior distribution of the small area totals Y_i or $g(Y_i) = \theta_i$ is then obtained. Inferences are based on the posterior distribution; in particular, Y_i or θ_i is estimated by its posterior mean and its precision is measured by its posterior variance.

For the basic area level model, (2.1) and (2.2), with normality of v_i and e_i, the posterior mean $E(\theta_i \mid \hat{\theta})$ and the posterior variance $V(\theta_i \mid \hat{\theta})$ are obtained in two stages, where $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)'$. In the first stage, we obtain $E(\theta_i \mid \hat{\theta}, \sigma_v^2)$ and $V(\theta_i \mid \hat{\theta}, \sigma_v^2)$ for fixed σ_v^2, assuming an improper prior, $f(\beta) \propto$ const., on β to reflect absence of prior information on β. The conditional posterior mean, given σ_v^2, is identical to the BLUP estimator $\tilde{\theta}_i(\sigma_v^2)$ and the conditional posterior variance is equal to $g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)$. At the second stage, we take account of the uncertainty about σ_v^2 by first calculating its posterior distribution $f(\sigma_v^2 \mid \hat{\theta})$, assuming a prior distribution on σ_v^2 and prior independence of β and σ_v^2. The posterior mean and variance are then obtained as

$$\tilde{\theta}_i^{\text{HB}} = E(\theta_i \mid \hat{\theta}) = E_{\sigma_v^2 \mid \hat{\theta}}\left[\tilde{\theta}_i(\sigma_v^2)\right] \tag{3.9}$$

$$V(\theta_i \mid \hat{\theta}) = E_{\sigma_v^2 \mid \hat{\theta}}\left[g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)\right] + V_{\sigma_v^2 \mid \hat{\theta}}[\bar{\theta}_i(\sigma_v^2)] \tag{3.10}$$

where $E_{\sigma_v^2|\hat\theta}$ and $V_{\sigma_v^2|\hat\theta}$ denote the expectation and variance with respect to $f(\sigma_v^2 \mid \hat\theta)$. No closed form expressions for (3.9) and (3.10) exist, but in this simple case they can be evaluated numerically using only one-dimensional integration. For complex models, high-dimensional integration is often involved and it is necessary to use MCMC-type methods to overcome the computational difficulties.

It follows from (3.9) that $\tilde\theta_i^{\text{HB}} \approx \tilde\theta_i(\hat\sigma_v^2)$ but (3.10) shows that ignoring uncertainty about σ_v^2 and using $g_{1i}(\hat\sigma_v^2) + g_{2i}(\hat\sigma_v^2)$ as a measure of variability can lead to significant underestimation.

If the assumed prior $f(\sigma_v^2)$ is proper and informative, the HB approach encounters no difficulties. On the other hand, an improper prior $f(\sigma_v^2)$ could lead to an improper posterior (Hobert and Casella 1996). In the latter case, we cannot avoid the difficulty by choosing a diffuse proper prior on σ_v^2 because we will be simply approximating an improper posterior by a proper posterior.

To illustrate the use of Gibbs sampling, we again consider the basic area level model under normality. To implement Gibbs sampling assuming the prior $f(\tau_v = \sigma_v^{-2})$ is a gamma (a, b), $a > 0$, $b > 0$, we need the following Gibbs-conditional distributions:

$$\text{(i)} \quad \beta \mid \theta, \sigma_v^2, \hat\theta \sim N_p[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\theta, \sigma_v^2(\mathbf{X}'\mathbf{X})^{-1}] \tag{3.11}$$

$$\text{(ii)} \quad \begin{gathered}\theta_i \mid \beta, \sigma_v^2, \hat\theta \overset{\text{ind}}{\sim} N(\tilde\theta_i(\beta, \sigma_v^2) = \\ \gamma_i\theta_i + (1-\gamma_i)\mathbf{x}_i'\beta, \gamma_i\psi_i), i = 1, \ldots, m\end{gathered} \tag{3.12}$$

$$\text{(iii)} \quad \sigma_v^{-2} \mid \beta, \theta, \hat\theta \sim G\left[\frac{m}{2} + a, \frac{1}{2}\sum(\theta_i - \mathbf{x}_i'\beta)^2 + b\right], \tag{3.13}$$

where $\mathbf{X}$ is the $m \times p$ matrix with $\mathbf{x}_i'$ as the i-th row and $\theta = (\theta_1, \ldots, \theta_m)'$. The Gibbs algorithm is as follows: (a) Using starting values $\theta_i^{(0)}$ and $\sigma_v^{2(0)}$ draw $\beta^{(1)}$ from (3.11). (b) Draw $\theta_i^{(1)}$, $i = 1, \ldots, m$ from (3.12) using $\beta^{(1)}$ and $\sigma_v^{2(0)}$. (c) Draw $\sigma_v^{2(1)}$ from (3.13) using $\theta_i^{(1)}$, $i = 1, \ldots, m$ and $\beta^{(1)}$. Steps (a)-(c) complete one cycle. Perform a large number of cycles, say t, called "burn-in period", until convergence and then treat $(\beta^{(t+j)}, \sigma_v^{2(t+j)}, \theta_i^{(t+j)}, j = 1, \ldots, J)$ as J samples from the joint posterior of β, σ_v^2 and θ_i, $i = 1, \ldots, m$. Other methods use multiple parallel runs instead of a single long run as above. Parallel runs can be wasteful because initial "burn-in" periods are discarded from each run. But a single long run may leave a significant portion of the space generated by the joint posterior unexplored.

The posterior mean and the posterior variance are estimated as

$$\tilde\theta_i^{\text{HB}} \approx \frac{1}{J}\sum_j \tilde\theta_i\left[\sigma_v^{2(t+j)}\right] = \frac{1}{J}\sum_j \tilde\theta_i(j) = \tilde\theta_i(\cdot) \tag{3.14}$$

and

$$V(\theta_i|\hat\theta) \approx \frac{1}{J}\sum_j \left[g_{1i}(\sigma_v^{2(t+j)}) + g_{2i}(\sigma_v^{2(t+j)})\right]$$
$$+ \frac{1}{J}\sum_j \left[\tilde\theta_i(j) - \tilde\theta_i(\cdot)\right]^2. \tag{3.15}$$

The estimator $\tilde\theta_i(\cdot)$ has smaller simulation error than the estimator is a conditional expectation and the well-known $J^{-1}\sum_j \theta_i^{(t+j)}$ because $\tilde\theta_i(\sigma_v^2)$ is a conditional expectation and the well-known Rao-Blackwell theorem holds. It is therefore advisable to do analytical calculations first before applying Gibbs sampling.

For the basic area level model, all the conditional distributions, (3.11)-(3.13), are in a closed form and, therefore, samples can be generated directly. But for more complex models, some of the conditionals may not have closed form in which case alternative algorithms, such as Metropolis-Hastings within Gibbs, are needed to draw samples from the joint posterior distribution. We refer the reader to Brooks (1998) for an excellent review of the MCMC methods. Software, called BUGS and CODA, are readily available for implementing MCMC and convergence diagnostics, but caution should be exercised in using MCMC methods. For example, Hobert and Casella (1996) demonstrated that the Gibbs sampler could lead to seemingly reasonable inferences about a nonexistent posterior distribution. This happens when the posterior is improper and yet all the Gibbs-conditional distributions are proper. Another difficulty with MCMC is that the convergence diagnostics tools can fail to detect the sorts of convergence failure that they were designed to identify (Cowles and Carlin 1996). Further difficulties include the choices of t for the burn-in period, number of simulated samples, J, and the starting values.

3.4 Some Recent Applications

(1) Dick (1995) used the basic area level model (2.3) to estimate net under coverage rates in the 1991 Canadian Census. The goal is to estimate 96 adjustment factors $\theta_i = T_i / C_i$, corresponding to 2(sex) × 4(age) × 12(province) combinations, where T_i is the true (unknown) count and C_i is the census count in the i-th area domain the net undercoverage rate in the i-th area is given by $U_i = 1 - \theta_i^{-1}$. Direct estimates $\hat\theta_i$ were obtained from a post enumeration survey, and sampling variances ψ_i were derived through smoothing of estimated variances, assuming ψ_i is proportional to some power of C_i. Explanatory variables, $\mathbf{x}$, were selected from a set of 42 variables by backward stepwise regression. EBLUP (EB) estimates of θ_i were used and their MSE estimated using (3.7) with REML estimate of σ_v^2. The EB adjustment factors $\tilde\theta_i^{\text{HB}}$ were converted to estimates of missing persons, $M_i = T_i - C_i$, and these estimates were raked to ensure consistency with direct estimates of marginal totals. The raked EB estimates, $\tilde\theta_i^{\text{R}}$ were used as the final

estimates of M_i's. MSE estimate of $\tilde{\theta}_i^{R}$ was obtained as $[\text{mse}(\tilde{\theta}_i^{HB})](\tilde{\theta}_i^{R}/\tilde{\theta}_i^{HB})^2$. This somewhat *ad hoc* method ensures that the coefficient of variation (CV) of $\tilde{\theta}_i^{HB}$ is retained by $\tilde{\theta}_i^{R}$, but properties of this method remains to be investigated.

(2) The basic area level model (2.3) with $\theta_i = \log Y_i$ has been recently used to produce model-based county estimates of poor school-age children in U.S.A. (Fisher and Siegel 1997; National Research Council 1998). Using these estimates, the US Department of Education allocates over 7 billion of federal funds annually to counties. The difficulty with unknown ψ_i was handled by using a model of the form (2.3) for the census year 1990, for which reliable estimates $\hat{\psi}_{ic}$ of sampling variances, ψ_{ic}, are available and assuming the census small area effects v_{ic} follow the same distribution as v_i, *i.e.*, $N(0, \sigma_v^2)$. Under the latter assumption, an estimate of σ_v^2 was obtained from the census data assuming $\hat{\psi}_{ic} = \psi_{ic}$ and used in the current model (2.3), assuming $\psi_i = \sigma_e^2/n_i$, to get an estimate of σ_e^2. The resulting estimate, $\tilde{\psi}_i = \tilde{\sigma}_e^2/n_i$, was treated as the true ψ_i in developing EBLUP estimates, $\tilde{\theta}_i$, of θ_i. The small area (county) totals Y_i (number of school-age children in poverty) can then be estimated as $\tilde{Y}_i = \exp(\tilde{\theta}_i)$, but a more refined method based on the mean of lognormal distribution was used: $\tilde{\tilde{Y}}_i = \exp\{\tilde{\theta}_i + ½\,\text{MSE}(\tilde{\theta}_i)\}$, ignoring the g_{3i}–term in (3.7) which was found to be small. The MSE of $\tilde{\tilde{Y}}_i$ was estimated using the approximation $\text{MSE}(\tilde{\theta}_i) \approx \text{CV}^2(\tilde{\tilde{Y}}_i)$. The estimates $\tilde{\tilde{Y}}_i$ were raked to agree with model-based state estimates obtained from a state model. The reader is referred to National Research Council (1998) for details on x-variables used in the county model and evaluation of the models. Several criteria were used for evaluating the models and the estimates, including regression diagnostics and comparisons to the 1990 Census counts.

(3) Other applications of the basic area level model include the following: (i) Estimation of unemployment rates at census tract level (Chand and Alexander 1995); (ii) Estimation of counts in employment categories and household income categories at the Congressional District level (Griffiths 1996); (iii) Estimation at the provincial level in the Italian Labour Force Survey (Falorsi, Falorsi and Russo 1995).

4. EXTENSIONS

We now present some recent extensions and applications of the basic area level model in section 4.1 and those of the basic unit level model in section 4.2.

4.1 Area-Level Models

Recent extensions of the basic area level model include multivariate and time series models and models for disease mapping, as noted in section 2.

4.1.1 Multivariate Models

Datta, Ghosh, Nangia and Natarajan (1996) used multivariate area level (Fay-Herriot) models to develop HB estimators of median income of four-person families for U.S. states. Here $\theta_i = (\theta_{i1}, \theta_{i2}, \theta_{i3})'$ with θ_{i1}, θ_{i2} and θ_{i3} denoting the true median incomes of four-, three- and five-person families in state i. Adjusted census median income and base- year census median income for the three groups were used as explanatory variables. Diffuse priors on model parameters were used along with Gibbs sampling. The resulting HB estimators, HB^3, for four-person families in 1979 were compared to the direct Current Population Survey (CPS) estimators and univariate and bivariate model-based HB estimators, HB^1 and HB^2, treating the 1979 estimates, available from the 1980 census data, as the true values. In terms of relative absolute error averaged over the states, the three HB estimators performed similarly, but outperformed the direct CPS estimates. In this application, the univariate estimator HB^1 worked well and it is not necessary to use more complicated estimators based on multivariate models. Estimates of θ_{i1} are used for administering an energy assistance program to low-income families.

Longford (1999) obtained multivariate shrinkage (composite) estimators of small area means and proportions, and illustrated their superiority over univariate shrinkage estimators.

4.1.2 Time Series Models

(1) Ghosh *et al.* (1996) developed HB estimators under the time series linking model given by (2.5) and (2.6) and applied them to estimate median income of four person families using direct estimates $\hat{\theta}_{it}, i = 1, \ldots, 51; t = 1, \ldots, 10$ for the 51 states over a ten year period.

(2) Datta *et al.* (1994) used the time series model (2.10) with $\mathbf{u}_t$ following (2.6) and developed HB estimators. They also used methods for validating the model, based on cross-validation. They applied the methods to estimate monthly unemployment rates for U.S. states. HB estimators performed significantly better than the CPS estimates, as measured by the CPS and HB standard errors. We refer the reader to Datta *et al.* (1994) for details on the x-variables used. Datta, Lahiri, Maiti, and Lu (1999) used the linking model (2.8) with a random walk model on the u_{it}'s, but added extra terms to (2.8) to reflect seasonal variation in unemployment rates.

(3) Datta, Lahiri and Maiti (1999) and You (1999) obtained EBLUP (EB) estimators and associated second-order correct estimators of MSE for the time

series/cross-sectional linking model (2.8) with a random walk model on u_{it}'s. Datta *et al.* used ML and REML estimators of model parameters while You employed the method of moments estimators.

Datta, Lahiri and Maiti (1999) used EB estimators to estimate median income of four-person families by U.S. states using time series and cross-sectional data. They employed the linking model (2.8) with a random walk model on u_{it}'s. Using the 1979 estimates available from the 1980 Census data as the true values, they compared the EB (EBLUP) estimates with the HB estimates of Ghosh *et al.* (1996) and the CPS direct estimates. In terms of absolute relative bias averaged over states, EB performed better than HB and both EB and HB performed much better than the CPS direct estimate. In terms of coefficient of variation, EB again performed better than HB and CPS; second-order correct estimate of MSE of EB was used.

4.1.3 Disease Mapping Models

Maiti (1998) used the model $y_i \mid \theta_i \overset{\text{ind}}{\sim} P(n_i\theta_i)$ and $\beta_i = \log\theta_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with diffuse prior on μ and a gamma prior on σ^{-2}. He obtained HB estimators of θ_i, and the posterior variance of θ_i, and applied the results to the well-known lip cancer data from Scottish Counties (small areas); see Clayton and Kaldor (1987) for details. He also studied HB estimation under the spatial dependence model for β_i's mentioned in section 2.1. Estimates of θ_i's are very similar for both the models but standard errors for the spatial model are smaller than those under the first model. Lahiri and Maiti (1996) obtained EB estimators and second order correct estimators of MSE under the Clayton-Kaldor model mentioned in section 2.1, and illustrated the method on the Clayton-Kaldor data set. Nandram *et al.* (1998) used the age-group specific models, mentioned in section 2.1, to obtain HB estimators and also developed Bayesian methods to compare alternative models, using three different measures of fit. They applied the results to estimate age specific and age adjusted mortality rates for Health Services Area's (sets of counties based on where residents seek routine hospital care) for the disease category "all cancers for white males".

4.1.4 Other Extensions

Datta and Lahiri (1995) considered robust HB estimation using a class of scale mixtures of normal distributions on the random effects v_i with the basic area level model. This class includes *t*, Laplace and logistic distributions; Cauchy distribution for outlier areas was adopted.

You (1999) considered the more realistic sampling model (2.4) on $\hat{Y}_i$ with sampling errors e_i^* and the linking model (2.1). Assuming $V(e_i^* \mid Y_i) = \psi_i^2 Y_i^2$ and $\hat{\theta}_i = \log(\hat{Y}_i)$, he used HB methods to demonstrate that for small sample sizes the posterior inferences under the sampling model (2.4) can be significantly different from those under the sampling model on $\hat{\theta}_i$.

4.2 Unit Level Models

Recent extensions at the basic unit level model include multivariate models, two-way and two-level models, random error variance models and logistic linear mixed models, as noted in section 2.

4.2.1 Nested Error Regression Models

Rao and Choudhry (1995) provided an overview of small area estimation in the context of business surveys. They also studied the performance of EBLUP estimator of a small area total relative to traditional estimators through simulation using real and synthetic populations.

As noted in section 2, model-based estimators for unit level models do not depend on the survey weights. Prasad and Rao (1999) obtained model-assisted estimators for the nested error regression model that depend on survey weights $\tilde{w}_{ij}$ and remain design-consistent as the sample size, n_i, increases. The unit level sample model is first reduced to

$$\bar{y}_{iw} = \bar{\mathbf{x}}_{iw}'\boldsymbol{\beta} + v_i + \bar{e}_{iw}, \tag{4.1}$$

where $\bar{y}_{iw} = \sum_j w_{ij} y_{ij}$ with $w_{ij} = \tilde{w}_{ij} / \sum_j \tilde{w}_{ij}$ and similar expressions for $\bar{x}_{iw}$ and $\bar{e}_{iw}$. A pseudo-BLUP estimator of $\theta_i = \bar{\mathbf{X}}_i'\beta + v_i$, for fixed σ_v^2 and σ_e^2, say $\hat{\theta}_{iw}(\sigma_v^2, \sigma_e^2)$ is then obtained from the reduced model (4.1), noting that $\bar{e}_{iw} \overset{\text{ind}}{\sim} N(0, \sigma_e^2 \sum_j w_{ij}^2)$, where $\bar{\mathbf{X}}_i$ is the vector of known population means and $\bar{Y}_i \approx \theta_i$ for large N_i (This estimator is called pseudo-BLUP because it is different from the BLUP estimator under the full unit-level sampling model). The unknown parameters σ_v^2 and σ_e^2 are then replaced by model-consistent estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ under the full model to obtain the pseudo-EBLUP estimator $\hat{\theta}_{iw} = \hat{\theta}_{iw}(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. This estimator is model-assisted and it is approximately design and model unbiased even if the sample design is nonignorable. Prasad and Rao (1999) also obtained a second-order correct estimator of $\text{MSE}(\hat{\theta}_{iw})$. You and Rao (1999b) developed a pseudo-HB methodology which leads to estimators similar to the pseudo-EBLUP estimators of Prasad and Rao (1999).

Singh, Stukel and Pfeffermann (1998) made a comparison of frequentist and Bayesian measures of error, using analytical and empirical methods for the basic unit-level model.

Stukel and Rao (1999) obtained EBLUP estimators and associated approximately unbiased (or second-order) correct MSE estimators under two-way nested error regression models. Simulation results of Stukel and Rao (1999) suggested that the behaviour of relative bias of MSE estimators is more complex than in the one-way case.

4.2.2 Two-Level Models

Moura and Holt (1999) obtained EBLUP estimators and associated second-order correct MSE estimators for the two-level models. They obtained EBLUP estimators and used them on data from a sample of 951 retail stores in Southern Brazil classified into 73 small areas. They compared the average second order correct MSE of the estimators to the average MSE value for the nested error regression model to demonstrate improvement in efficiency. You and Rao (1999a) applied HB methods to the Brazilian data. They studied three different two level models: (1) equal error variances; (2) unequal error variances; (3) random error variances. Bayesian diagnostics revealed that model (2) fits the data better than models (1) and (3).

4.2.3 Random on Error Variances Models

Arora and Lahiri (1997) studied the unit-level model with random error variances σ_i^2 and assumed $\sigma_i^{-2} \overset{\text{i.i.d.}}{\sim} G(a, b)$. They obtained the EB estimator of small area mean $\bar{Y}_i$ and applied the Laird-Louis bootstrap to estimate its MSE, taking account of the variability due to estimation of model parameters.

Arora and Lahiri (1997) obtained a reduced model from the unit level random error variances model by incorporating survey weights. They performed HB analysis on the reduced model with $\sigma_i^{-2} \overset{\text{i.i.d.}}{\sim} G(a, b)$, and applied the results to estimate the average weekly consumer expenditures of various items, goods and services for $m = 43$ publication areas (small areas) in U.S.A.

4.2.4 General Linear Mixed Models

Datta and Lahiri (1997) studied the general linear mixed model with a block diagonal covariance structure, (2.12). They developed EBLUP estimators and associated second-order correct estimators of MSE, using REML or ML estimators. In the case of ML estimators an extra term of order $O(m^{-1})$ should be subtracted. Das, Rao et You (1999) extended these results to the general mixed ANOVA model (2.13) in which case the asymptotic set-up is more complex.

4.2.5 Multivariate Nested Error Regression Models

Datta, Day and Basawa (1999) obtained EBLUP (EB) estimators and second order correct estimators of MSE, for the multivariate nested error regression models. They conducted a simulation study using the sample sizes and auxiliary variable values given by Battese, Harter and Fuller (1988). Further, they estimated the model parameters for their multivariate model using Battese *et al.*, data on crop areas under corn and soybeans for $m = 12$ counties in North-Central Iowa. Treating the estimated parameters as true values, they generated simulated samples and showed that the multivariate approach can achieve substantial improvement over the univariate approach inefficiency.

4.2.6 Logistic Linear Mixed Models

Farrell, MacGibbon and Tomberlin (1997a, 1997b) studied EB estimation for binary y, assuming the sampling model $y_{ij} \mid \theta_{ij} \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{ij})$ and the linking logistic model $\log\{\theta_{ij}/(1-\theta_{ij})\} = \mathbf{x}_{ij}'\beta + v_i$ with $v_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_v^2)$. The conditional distribution of θ_{ij}'s is approximated by a multivariate normal to get an EB estimator of local area proportion $\bar{Y}_i$. They employed the bootstrap method of Laird and Louis (1987) to get a bootstrap-adjusted estimate of variability associated with the EB estimator. But results of Butar and Lahiri (1997) for the linear case suggest that the bootstrap method may not be second-order correct in the nonlinear case as well. Jiang and Lahiri (1998) also studied EB estimation for the above model and obtained the EB estimator exactly through one-dimensional numerical integration. They called the EB estimator an empirical best predictor (EBP) which may be more appropriate because no priors on model parameters are involved. Employing method of moment estimators of model parameters β and σ_v^2, they also obtained an approximation to MSE of the EB estimator correct to terms of order m^{-1}. Jiang, Lahiri, and Wan (1999) proposed a jackknife method of estimating MSE that is applicable to general longitudinal linear and generalized linear mixed models. This method leads to second-order correct MSE estimators and looks promising. But one needs to recompute the REML estimates of model parameters by deleting each area in turn. The computations can be significantly reduced by using a single step of the Newton-Raphson algorithm with the estimates from the full sample as starting values. Properties of this simplification remain to be studied. Booth and Hobert (1998) argued that the conditional MSE of the EBP given the i-th area data is more relevant as a measure of variability than the unconditional MSE because it is area-specific. Fuller (1989) earlier proposed a similar criterion in the context of linear mixed models. But the MSE estimator (3.8) shows that it is possible to obtain area-specific estimators of the unconditional MSE, at least in the linear model case. Also, it is not clear how one should proceed with the conditioning when two or more small area estimators need to be aggregated to obtain an estimator for a larger area. How would one define the conditional MSE of the larger area estimator?

Malec *et al.* (1997) used logistic linear mixed models and the HB approach to estimate proportions for demographic groups within U.S. states. Data from the National Health Interview Survey were used for this purpose. Cross-validation methods were used to evaluate the model fit. For one of the binary variables observed for respondents to the 1990 census long form, they compared the estimates from alternative methods and models with the very accurate census estimates of true values. For logistic linear mixed models, not all the conditional distributions for Gibbs sampling have closed form unlike those obtained for the probit linear mixed model derived from a latent variable approach (Das *et al.* 1999).

Malec, Davis and Cao (1996, 1999) studied logistic linear mixed models to estimate overweight prevalence for subgroups (small areas) using National Health and Nutrition Examination Survey (NHANES III) data. Again, HB methods were used but survey weights were incorporated using a pseudo-likelihood. Folsom, Shah and Vaish (1999) studied general logistic mixed linear models in the context of estimating substance abuse in U.S. states from the 1994-1996 National Household Surveys on Drug Abuse. They developed survey-weighted pseudo HB estimators and associated posterior variance, using MCMC methods.

Ghosh *et al.* (1998) applied the HB approach to generalized linear mixed models and used the results on two real data sets. The first data set, based on a 1991 sample of all persons in 15 geographical regions of Canada consists of responses classified into four categories to the question "Have you experienced any negative impact of exposure to health hazards in the work place?" Objective here is to estimate the proportion of workers in each of the four response categories for every one of 60 groups cross-classified by 16 geographical regions and 4 demographic (age × sex) groups. The second data set relates to cancer mortality rates for the 115 counties in Missouri during 1972-81.

5. DISCUSSION

We briefly discussed, in section 1, survey design issues that have an impact on small area statistics. Preventive measures at the design stage, such as those proposed by Singh, Gambino and Mantel (1994), may reduce the need for indirect estimators significantly, although for many applications sample sizes in some domains of interest may not be large enough to provide adequate precision even after taking such measures. As noted in section 1, sometimes the survey is deliberately designed to oversample specific domains at the expense of small samples or even no samples in other domains (areas) of interest.

We have provided a brief overview of the literature, over the past five years or so, on model-based small area estimation. The methodological developments and applications are both impressive, but it is necessary to exercise caution in using model-based methods because of the underlying assumptions. Good auxiliary information related to the variables of interest plays a vital role in model-based inference. As noted by Schaible (1996), expanded access to auxiliary information through coordination and cooperation among federal agencies is needed.

Model validation also plays an important role in model-based estimation. Fay and Herriot (1979), Ghosh and Rao (1994), Dick (1995), Malec *et al.* (1997), Datta, Lahiri, Maiti, and Lu (1999), You and Rao (1999a), National Research Council (1998) and others used some methods for model validation and illustrated their application. But the available methods for handling models with random effects are not as extensive as those used for the standard linear and non-linear models with only fixed effects. More work, both classical and Bayesian, on model diagnostics for random effects models is needed.

Area-level models have wider scope than the unit level models because area-level auxiliary information is more readily available than unit-level auxiliary data. But the assumption of known sampling variances, ψ_i, is quite restrictive, although the methods used in the applications (section 3.4) seem to be promising. It should be noted that errors in estimating ψ_i do not affect the model-unbiasedness of the EBLUP(EB) estimators provided the mean of θ_i in the linking model (2.1) is correctly specified. But the efficiency of the estimator is affected as well as the validity of the MSE estimators (3.7) and (3.8). More work is needed on obtaining good approximations to the sampling variances. This task becomes more difficult when using multivariate and time series area levels models because sampling covariances are also needed.

Recent work on incorporating survey weights into model-based estimation under unit-level models is promising, but the assumption that the sample design is ignorable may not be true for some applications. Krieger and Pfeffermann (1997) proposed methods for direct estimation of large area parameters that take account of the sample selection effects. It would be useful to extend this work to indirect estimation of small area parameters.

The Hierarchical Bayes (HB) approach is a powerful method for small area estimation because it can handle complex problems and the inferences are "exact". But, as noted in section 3.3, caution should be exercised in the choice of improper prior distributions on the model parameters.

We studied model-based estimates of small area totals or means, but they may not be suitable if the objective is to identify domains with extreme population values or to rank domains or to identify domains that fall below or above some prespecified level. Ghosh and Rao (1994) reviewed some methods for handling the latter cases. For a simple model, Shen and Louis (1998) proposed "triple-goal" estimators that produce good ranks, a good distribution and good area – specific estimators. It would be useful to extend their approach to handle more complex models that are suitable for small area estimation.

ACKNOWLEDGEMENTS

This work was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada. I am thankful to Dr. Graham Kalton and the Editor for constructive suggestions.

REFERENCES

ARORA, V., and LAHIRI, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

BATTESE, G.E., HARTER, R.M., and FULLER, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

BOOTH, J.G., and HOBERT, J.P. (1998). Standard errors of predictors in generalized linear mixed models. *Journal of the American Statistical Association*, 93, 362-372.

BROOKS, S.P. (1998). Markov Chain Monte Carlo method and its application. *The Statistician*, 47, 69-100.

BUTAR, P.B., and LAHIRI, P. (1997). On the Measures of Uncertainty of Empirical Bayes Small Area Estimators. Technical Report, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

CHAND, N., and ALEXANDER, C.H. (1995). Indirect estimation of rates and proportions for small areas with continuous measurement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 549-554.

CHAUDHURI, A., and ADHIKARY, A.K. (1995). On generalized regression estimators of small domain totals – an evaluation study. *Pakistan Journal of Statistics*, 11, 173-189.

CLAYTON, D., and KALDOR, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

COWLES, M.K., and CARLIN, B.P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91, 883-904.

DAS, K., and RAO, J.N.K. (1999). Second order approximations for standard errors of empirical BLUP estimators in general mixed ANOVA models. Paper under preparation.

DAS, K., RAO, J.N.K., and YOU, Y. (1999). Small Area Estimation for Binary Variables Using Probit Linear Mixed Models. Paper under preparation.

DATTA, G.S., DAY, B., and BASAWA, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference*, 75, 269-279.

DATTA, G.S., GHOSH, M., NANGIA, N., and NATARAJAN, K. (1996). Estimation of median income of four-person families: a Bayesian approach. In: *Bayesian Analysis in Statistics and Econometrics*, (D.A. Berry, K.M. Chaloner, and J.K. Geweke, Eds.). New York: Wiley, 129-140.

DATTA, G.S., and LAHIRI, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis*, 54, 310-328.

DATTA, G.S., and LAHIRI, P. (1997). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictor in Small-Area Estimation Problems. Technical Report, Department of Statistics, University of Georgia-Athens.

DATTA, G.S., LAHIRI, P., and LU, K.L. (1994). Hierarchical Bayes Time Series Modeling in Small Area Estimation With Applications. Technical Report, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

DATTA, G.S., LAHIRI, P., and MAITI, T. (1999). Empirical Bayes Estimation of Median Income of Four Person Families by States Using Time Series and Cross-Sectional Data. Technical Report, Department of Statistics, University of Georgia-Athens.

DATTA, G.S., LAHIRI, P., MAITI, T., and LU, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the U.S. states. *Journal of the American Statistical Association*, 94, 1074-1082.

DICK, P. (1995). Modelling net undercoverage in the 1991 Canadian Census. *Survey Methodology*, 21, 45-54.

FALORSI, P.P., FALORSI, S. and RUSSO, A. (1994). Empirical comparison of small area estimation methods for the Italian labour force survey. *Survey Methodology*, 20, 171-176.

FALORSI, P.D., FALORSI, S. and RUSSO, A. (1995). Small area estimation at provincial level in the Italian Labour Force Survey. *Proceedings of the 1995 Annual Research Conference, U.S. Bureau of the Census*, 617-635.

FARRELL, P.J., MacGIBBON, B., and TOMBERLIN, T.J. (1997a). Empirical Bayes estimates of small area proportions in multistage designs. *Statistica Sinica*, 7, 1065-1083.

FARRELL, P.J., MacGIBBON, B., and TOMBERLIN, T.J. (1997b). Empirical Bayes small area estimation using logistic regression models and summary statistics. *Journal of Business and Economic Statistics*, 15, 101-108.

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.

FISHER, R., and SIEGEL, P. (1997). Methods used for small area poverty and income estimation. *Proceedings of the Social Statistics Section, American Statistical Association.*

FOLSOM, R., SHAH, B., and VAISH, A. (1999). Substance Abuse in States: Model Based Estimates from the 1994-1996 National Household Surveys on Drug Abuse. Methodology report. Research Triangle Institute.

FULLER, W.A. (1989). Prediction of True Values for the Measurement Error Model. Paper presented at the Conference on Statistical Analysis of Measurement Error Models, Humboldt State University.

FULLER, W.A., and HARTER, R.M. (1987). The multivariate components of variance model for small area estimation. In: *Small Area Statistics* (R. Platek, J.N.K. Rao, C.E. Särndal and M.P. Singh Eds.). New York: John Wiley, 103-123.

GHOSH, M. NANGIA, N., and KIM, D.H. (1996). Estimation of median income of four-person families: a Bayesian approach. *Journal of the American Statistical Association*, 91, 1423-1431.

GHOSH, M., NATARAJAN, K., KIM, D., and WALKER, L.A. (1997). Hierarchical Bayes GLM's for the Analysis of Spatial Data: An Application to Disease Mapping. Technical report, University of Florida-Gainsville.

GHOSH, M., NATARAJAN, K., STROUD, T.W.F., and CARLIN, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 55-93.

GRIFFITHS, R. (1996). Current Population Survey Small Area Estimation for Congressional Districts. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 314-319.

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, 1461-1479.

JIANG, J. (1996). REML estimation: asymptotic behaviour and related topics. *Annals of Statistics*, 24, 255-286.

JIANG, J., and LAHIRI, P. (1998). Empirical Best Prediction for Small Area Inference With Binary Data. Technical report, Department of Mathematics and Statistics, University of Nebraska- Lincoln.

JIANG, P., LAHIRI, P., and WAN, S. (1999). Jackknifing the Mean Squared Error of Empirical Best Predictor. Technical report, Department of Statistics, Case Western Reserve University.

JIANG, J., LAHIRI, P., and WU, C. (1998). On Pearson-χ^2Testing With Unobservable Frequencies and Mixed Model Diagnostics. Technical report, Department of Statistics, Case Western Reserve University.

KLEFFE, J., and RAO, J.N.K. (1992). Estimation of mean square error of empirical best linear unbiased predictors under random error variance linear model. *Journal of Multivariate Analysis,* 43, 1-15.

KRIEGER, A.M., and PFEFFERMANN, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics,* 13, 123-142.

LAHIRI, P.A., and MAITI, T. (1996). Empirical Bayes Estimation of Mortality From Diseases for Small Area. Technical report, Department of Mathematics and Statistics, University of Nebraska-Lincoln.

LAHIRI, P.A., and RAO, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association,* 82, 758-766.

LAIRD, N.M., and LOUIS, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association,* 82, 739-750.

LONGFORD, N.T. (1999). Multivariate shrinkage estimations small area means and proportions. *Journal of the Royal Statistical Society,* Series A, 182, 227-245.

MAITI, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference,* 69, 339-348.

MALEC, D., DAVIS, W.W., and CAO, X. (1996). Small area estimates overweight prevalence using the third National Health and Nutrition Examination Survey (NHANES III). *Proceedings of the Section on Survey Research Method, American Statistical Association,* 326-331.

MALEC, D., DAVIS, W.W., and CAO, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine,* 18, 3189-3200.

MALEC, D., SEDRANSK, J., MORIARITY, C.L., and LECLERE, F. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association,* 92, 815-826.

MARKER, D.A. (1999). Organization of small area estimators using generalized linear regression framework. *Journal of Official Statistics,* 15, 1-24.

MOURA, F., and HOLT, D. (1999). Small area estimation using multi level models. *Survey Methodology,* 25, 73-80.

NANDRAM, B., SEDRANSK, J., and RICKLE, L. (1998). Regression Analysis of Mortality Rates for U.S. Health Service Areas. Technical report, Worcester Polytechnic Institute.

NATIONAL RESEARCH COUNCIL (1998). *Small Area Estimation of School-Age Children in Poverty.* Interim Rep. 2. Washington, D.C.: National Research Council.

PRASAD, N.G.N., and RAO, J.N.K. (1990). The estimation of mean squared errors of small-area estimators. *Journal of the American Statistical Association,* 85, 163-171.

PRASAD, N.G.N., and RAO, J.N.K. (1999). On robust estimation using a simple random effects model. *Survey Methodology,* 25, 67-72.

RAGHUNATHAN, T.E. (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association,* 88, 1444-1448.

RAO, J.N.K. (1998). EB and EBLUP in Small Area Estimation. Technical report, Laboratory for Research in Statistics and Probability, Carleton University.

RAO, J.N.K., and CHOUDHRY, G.H. (1995). Small area estimation: overview and empirical study. In: *Business Survey Methods* (B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, Eds.). New York: John Wiley, 527-542.

RAO, J.N.K., and YU, M. (1992). Small area estimation by combining time series and cross-sectional data. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 1-9.

RAO, J.N.K., and YU, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics,* 22, 511-28.

RIVEST, L.P., and BELMONTE, E. (1999). The Conditional Mean Squared Errors of Small Area Estimators in Survey Sampling. Technical report, Laval University.

SCHAIBLE, W.L. (Editor) (1996). *Indirect Estimators in U.S. Federal Programs. Lecture Notes in Statistics No. 108.* New York: Springer-Verlog.

SHEN, W., and LOUIS, T.A. (1998). Triple-goal estimators in two-stage hierarchical models. *Journal of the Royal Statistical Society,* Series B, 60, 455-471.

SINGH, A.C., MANTEL, H.J., and THOMAS, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology,* 20, 33-43.

SINGH, A.C., STUKEL, D.M., and PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society,* Series B, 60, 377-396.

SINGH, M.P., GAMBINO, J., and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology,* 20, 3-22.

STUKEL, D.M., and RAO, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference,* 78, 131-147.

YOU, Y. (1999). Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation. Ph.D. Thesis, School of Mathematics and Statistics, Carleton University.

YOU, Y., and RAO, J.N.K. (1999a). Hierarchical Bayes estimation of small area means using multi-level models. Invited paper, *Proceedings of the IASS Satellite Conference on Small Area Estimation,* Riga, Latvia, 171-185.

YOU, Y., and RAO, J.N.K. (1999b). Pseudo hierarchical Bayes small area estimation using sampling weights. *1999 Proceedings of the Survey Methods Section, Statistical Society of Canada,* in press.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 187-191
Statistics Canada

Population Based Establishment Sample Surveys: The Horvitz-Thompson Estimator

MONROE SIRKEN and IRIS SHIMIZU[1]

ABSTRACT

The Population Based Establishment Survey (PBES) is a linked population/establishment sample survey in which listings of the establishments having transactions with households in a population sample survey serve as sampling frames for establishment surveys. This paper presents and discusses the PBES Horvitz-Thompson estimator of X, the sum of a variate over the transactions of all establishments.

KEY WORDS: Network sampling; Establishment transactions; Integrated sample design.

1. INTRODUCTION

Whenever free-standing sampling frames are unavailable or when available frames lack good coverage of establishments or lack good measures of establishment size, the Population Based Establishment Survey (PBES) is an attractive design alternative to the conventional establishment sample survey. And whenever the variate of interest refers to rare and elusive populations that are hard to reach directly, the PBES is an attractive design alternative to the conventional population sample survey.

This paper presents the PBES Horvitz-Thompson estimator of X, the sum of a variate over the M transactions of R establishments. Let M_j be the total number of transactions of the establishment $E_j (j = 1, ..., R)$ during a specified calendar period. The task at hand is to design a multipurpose establishment sample survey to estimate the Xs for a large number of different variates. Typically, establishment surveys that seek to estimate X are designed as two-stage sample surveys in which establishments are selected with probabilities proportionate to size, and their transactions are the second stage selection units. Designed in this manner, establishment surveys require free-standing sampling frames with good coverage of R establishments and good measures of establishment sizes, the M_js.

Though listings of households and persons enumerated in population sample surveys often serve as sampling frames for other population sample surveys (Mathiowetz 1987; Cox, Folsom and Virage 1987), listings of establishments that have transactions with households in population sample surveys rarely serve as frames for establishment sample surveys. The Consumer Price Index (CPI) which depends on data collected in population and establishment surveys (Leaver and Valliant 1995) is a notable exception. Households enumerated in the CPI Continuing Point of Purchase Survey (CPOPS), a population sample survey, report the establishments with whom they had transactions (purchased merchandise). The listing of establishments reported in CPOPS serves as the sampling frame for the CPI Pricing Survey, a sample survey of retail establishments that collects prices for a basket of consumer goods.

Several years ago, a Panel of the Committee on National Statistics, National Research Council (Wunderlich 1992), suggested that the National Center for Health Statistics (NCHS) investigate the feasibility and potential gains of using listings of medical providers that are reported by households in the National Health Interview Survey (NHIS) as sampling frames for NCHS' national medical provider sample surveys which were then and still are independently designed as conventional establishment sample surveys. [The NHIS is an on-going household survey of about 42,000 households annually that is conducted by the NCHS to obtain national health statistics for the U.S. civilian non-institutional population (Massey, Moore, Parsons and Tadros 1989)]. The Committee's suggestion initiated a PBES research program at NCHS.

Judkins, Berk, Edwards, Mohr, Stewart and Waksberg (1995) compared the operational and design features of the health care surveys if linked to NHIS with design features of independently designed health care surveys. Judkins, Marker, Waksberg, Botman and Massey (1999) made rough cost/error comparisons of an independently designed dental survey and a PBES dental survey linked to NHIS. They tentatively concluded that if a reasonable list with a reasonable measure of size can be found, an independently designed dental survey is probably preferable, and otherwise the dental survey linked to NHIS should be considered.

More recently, the PBES research has been theoretically oriented, focusing on the problem of constructing alternative unbiased PBES estimators with different data requirements, and getting closed formulas for their variances. Conceptual difficulties initially encountered in this effort were overcome once it was recognized that the PBES is a population network sample survey (Sirken 1970). Applying network sampling theory, Sirken, Shimizu, and Judkins (1995), and Shimizu and Sirken (1998) obtained two

[1] Monroe Sirken and Iris Shimizu, National Center for Health Statistics, 6525 Belcrest Road, Room 700, Hyattsville, MD 20782, U.S.A. e-mail: mgs2@cdc.gov.

versions of the unbiased PBES multiplicity estimator and derived their variances. In this paper, we present the unbiased PBES Horvitz-Thompson estimator and its variance. The PBES estimators are essentially extensions to multiple stage sampling under special conditions of single-stage network sampling estimators that were originally proposed by Birnbaum and Sirken (1965), and described by Thompson (1992).

2. NOTATION

Let M_j represent the number of transactions of the establishment $E_j (j = 1, ..., R)$. Then

$$M = \sum_{j=1}^{R} M_j = \text{the total number of transactions of the } R \text{ establishments.} \quad (1)$$

Let N_j = the number of households having transactions with establishment $E_j (j = 1, ..., R)$, N_{jl} = number of households having transactions with both establishments E_j and $E_l (j \neq l)$, and N_0 = number of households not having any transactions with any establishments. Then

$$N^* = \sum_{j=1}^{R} N_j - \sum \sum_{j \neq l} N_{jl} = \text{the total number of households having transactions with } R \text{ establishments,} \quad (2)$$

and

$$N = N^* + N_0 = \text{the total number of households.} \quad (3)$$

Let the value of the variate for the kth $(k = 1, ..., M_j)$ transaction of the establishment $E_j (j = 1, ..., R)$ be denoted by X_{jk}. Then

$$X_j = \sum_{k=1}^{M_j} X_{jk} = \text{sum of the variate over } M_j \text{ transactions of the establishment } E_j, \quad (4)$$

and

$$X = \sum_{j=1}^{R} X_j = \text{sum of the variate over } M \text{ transactions of } R \text{ establishments.} \quad (5)$$

3. THE NETWORK SAMPLING ERROR MODEL

A PBES is conducted to estimate X. First, a population sample survey based on a random sample of n households $H_i (i = 1, ..., n)$ is conducted in which sample households identify each of the establishments with whom they had transactions during a specified calendar period. After eliminating duplicate reports of the same establishments, a follow-on establishment survey is conducted with the r distinct establishments reported by n households in the population sample survey, and each sample establishment $E_j (j = 1, ..., r)$ independently selects and reports the variates for a random sample m_j of its M_j transactions.

Judkins *et al.* (1999) view the PBES as a 2-stage establishment sample survey in which the r establishments that had transactions with n sample households in the population survey are first stage selection units, and the m_j transactions $(j = 1, ..., r)$ selected by each of the r establishments, are second stage sampling units. However, the PBES design features become more transparent, and the PBES estimators and their variances more tractable when the PBES is modeled as a 2-stage network sample population survey. From the network sampling perspective, households are first stage units, and transactions that are countable at sample households in compliance with the PBES counting rule are second stage units.

The PBES counting rule specifies that every household in the network of N_j households that had transactions with $E_j (j = 1, ..., R)$ is linked to the same fixed size random sample m_j of the M_j transactions of the establishment E_j. The PBES counting rule implies that the same m_j transactions of $E_j (j = 1, ..., R)$ are countable in the population survey at every sample household belonging to the network of N_j households that had transactions with E_j. From the network sampling perspective, establishments that have transactions with households are proxy respondents for transactions that are countable at households. PBES households do not report about their own transactions nor about the transactions countable at their addresses vis-a-vis the PBES counting rule. Households identify establishments with whom they had transactions and those establishments select the subsamples of their transactions that are countable at sample households and they report the variates for the selected transactions.

The PBES counting rule produces a configuration of transactions between establishments and households that partitions the N households into R establishment networks, $A_j (j = 1, ..., R)$, where the network A_j contains the set of N_j households and is linked to the M_j transactions of E_j. Though the same household may belong to multiple networks, each of the M transactions is uniquely linked to one and only network.

Networks are counted differently by PBES multiplicity estimators and by the Horvitz-Thompson estimator. Multiplicity estimators count the M_j transactions linked to the network $A_j (j = 1, ..., R)$ every time households belonging to the network A_j are selected in the population survey sample. The Horvitz-Thompson estimator does not depend on the number of times that households belonging to the same networks are selected in the population survey. The PBES Horvitz-Thompson estimator counts each distinct network only once.

4. THE PBES HORVITZ-THOMPSON ESTIMATOR

For a sample of n households selected by simple random sampling, and a total sample of

$$m = \sum_{j=1}^{r} m_j = \text{transactions}, \tag{6}$$

where the transaction subsamples $m_j\,(j = 1, ..., r)$ are selected independently and by simple random sampling, the PBES Horvitz-Thompson estimator of X is

$$X' = \sum_{j=1}^{R} \frac{\alpha_j}{p_j} X_j'. \tag{7}$$

Here α_j is a random variable that is equal to 1 if any of the n sample households belongs to the network A_j and α_j is equal to 0 otherwise, and

$$X_j' = M_j \sum_{k=1}^{m_j} \frac{X_{jk}}{m_j}$$

is the unbiased estimator of $X_j\,(j = 1, ..., R)$ (8)

and

$p_j = E(\alpha_j) =$
the probability of any of the n sample households belonging to network $A_j (j = 1, ..., R)$. (9)

X' is an unbiased estimate of X if everyone of the R establishments has transactions with at least one household.

Let

$q_j = 1 - p_j =$
the probability that none of the n sample households belongs to the network A_j. (10)

If n households are selected by simple random sampling without replacement,

$$q_j = \frac{\binom{N - N_j}{n}}{\binom{N}{n}}. \tag{11}$$

If n households are selected by simple random sampling with replacement,

$$q_j = \frac{(N - N_j)^n}{N^n}. \tag{12}$$

There are two potential measurement problems involving the q_js $(j = 1, ..., r)$. First, they are dependent on the N_js $(j = 1, ..., r)$, quantities that are often difficult to ascertain in establishment surveys. Second, it would be difficult to compute the q_js for most population surveys which, like the NHIS, are based on complex sample designs.

5. THE VARIANCE OF THE PBES HORVITZ-THOMPSON PBES ESTIMATOR

The variance of the Horvitz-Thompson estimator of X may be written as

$$\text{Var}\,(X') = \text{Var}E(X'|\Omega) + E(\text{Var}\,X'|\Omega) \tag{13}$$

where $(X'|\Omega)$ denotes the value of X' conditional on a fixed sample Ω of n households.

Consider the first term on the right side of (9),

$$\begin{aligned} \text{Var}E(X'|\Omega) &= \text{Var}\left(\sum_{j=1}^{R} \frac{\alpha_j X_j}{p_j}\right) \\ &= \sum_{j=1}^{R} \frac{X_j^2}{p_j^2} \text{Var}(\alpha_j) \\ &\quad + \sum_{j=1}^{R} \sum_{l \neq j} \frac{X_j}{p_j} \frac{X_l}{p_l} \text{Cov}(\alpha_j \alpha_l). \end{aligned} \tag{14}$$

Since α_j is a binomial random variable

$$\text{Var}(\alpha_j) = p_j - p_j^2 \tag{15}$$

and

$$\text{Cov}(\alpha_j \alpha_l) = p_{jl} - p_j p_l \tag{16}$$

where

$$p_{jl} = 1 - q_j - q_l + q_{jl}^* \quad (j \neq l) \tag{17}$$

is the joint probability that any of the n sample households belongs to the networks A_j and A_l, and $q_{jl}^* (j \neq l)$ is the probability that the n sample households are linked to neither the network A_j nor A_l.

For simple random sampling of n households with replacement,

$$q_{jl}^* = \frac{(N - N_j - N_l + N_{jl})^n}{N^n}, \tag{18}$$

and for simple random sampling of n households without replacement,

$$q_{jl}^{*} = \frac{\binom{N-N_j-N_l+N_{jl}}{n}}{\binom{N}{n}}. \qquad (19)$$

Consider the second term on the right side of (13),

$$E(\text{Var}\,X'|\Omega) = E\left[\sum_{j=1}^{R} \frac{\alpha_j}{p_j^2} M_j^2 \,\text{Var}\,(\bar{X}_j')\right]$$

$$= \sum_{j=1}^{R} M_j^2 \frac{\text{Var}(\bar{X}_j')}{p_j}. \qquad (20)$$

$$\text{Var}\,(\bar{X}_j') = \frac{M_j - m_j}{m_j M_j}\,\sigma^2(X_j) \qquad (21)$$

where the population variance

$$\sigma^2(X_j) = \frac{\sum_{k=1}^{M_j} (X_{jk}-\bar{X}_j)^2}{M_j - 1}. \qquad (22)$$

Suppose the PBES is designed as a self-weighting sample. Then, $rp_j(m_j/M_j) = f$, where f is the overall sampling ratio of selecting a transaction. For a prescribed value of f, the size of the sample of transactions selected in the establishment E_j is

$$m_j = \frac{fM_j}{rP_j} = m\,\frac{M_j/p_j}{\sum_{j=1}^{r} M_j/M_j/p_j}\quad j = 1, \ldots, r. \qquad (23)$$

Combining (14) and (20), the variance of the PBES Horvitz-Thompson estimator of X is

$$\text{Var}(X') = \sum_{j=1}^{R} \frac{1-p_j}{p_j} X_j^2 + \sum_{j=1}^{R}\sum_{l \neq j} \frac{p_{jl} - p_j p_l}{p_j p_l} X_j X_l$$

$$+ \sum_{j=1}^{R} \frac{M_j^2}{p_j} \frac{M_j - m_j}{m_j M_j} \sigma^2(X_j). \qquad (24)$$

The first two terms on the right side of (24) represent the between establishment component of variance due to sampling households. The second term on the right vanishes if none of the N households has transactions with more than one establishment. The third term on the right side of (24) is the within establishment component of the variance due to subsampling transactions, and vanishes in single stage sampling when the sample establishments report the variates for all their transactions. Single stage sampling is more likely to be the design option in a single purpose PBES than in a multi-purpose PBES, especially when the variate of interest represents a relatively rare event.

6. CONCLUDING REMARKS

All unbiased PBES estimators, whether the Horvitz-Thompson estimator proposed in this paper or the PBES multiplicity estimators proposed by Sirken, Shimizu and Judkins (1995) and Shimizu and Sirken (1998) depend on multiplicity parameters to adjust for variations in the selection probabilities of the establishments reported in the population sample survey. However, multiplicity and Horvitz-Thompson estimators differ in the ways multiplicities are defined and in likelihood of successfully collecting this information in the follow-on survey with the establishments that were reported in the population survey.

The feasibility and ease with which establishments can provide the multiplicity information is a key factor in deciding on which kind of PBES estimator, if any, is most appropriate in particular applications. The N_js and M_js ($j = 1, \ldots, r$) respectively are the multiplicities needed by the PBES Horvitz-Thompson estimator and the PBES multiplicity estimators, where N_j is the number of households having transactions with the establishment E_j, and M_j is the total number of transactions of the establishment E_j. The N_js are unlikely to be readily available except at establishments, such as health maintenance organizations, utility companies, and home owner insurance companies, for which households are the transactional units. On the other hand, the M_js are likely to be available at many establishments that tend to keep track of the total number of services provided though unlikely to know the number of households to whom services were provided.

The PBES is a sample survey design option with many potential applications. It is a mechanism for linking population sample surveys to data files of establishments. Because the mechanism does not require disclosure of personal identifiers, PBES would not be restricted by the kinds of confidentiality concerns that ordinarily limit access to establishment data files. PBES offers the prospects of being able to estimate the volume of establishment transactions under circumstances beyond the capabilities of conventional establishment sample surveys when free-standing establishment frames are unavailable or inadequate, and beyond the capabilities of conventional population sample surveys when the variates of interest relate to rare and elusive populations that are hard to reach directly. Determining which, if any, of these and possibly other potential PBES contributions are realizable will require research studies comparing the cost and error effects of PBES estimators and estimators of conventional establishment and population sample surveys.

ACKNOWLEDGEMENTS

The authors thank the Panel of the National Health Care Survey of the Committee on National Statistics for suggesting the problem addressed in this paper, and also thank a referee for very helpful comments. Views presented in this paper are those of the authors and do not necessarily represent the official views or positions of the National Center for Health Statistics.

REFERENCES

BIRNBAUM, Z., and SIRKEN, M. (1965). Design of sample surveys to estimate the prevalence of rare diseases: three unbiased estimates. National Center for Health Statistics. *Vital and Health Statistics*, Series 2, No. 11. Washington, DC: Government Printing Office.

COX, G.B., FOLSOM, R.E., and VIRAGE, T.G. (1987). Design alternatives for integrating the National Medical Expenditure Survey with the National Health Interview Survey. National Center for Health Statistics. *Vital and Health Statistics*, Series 2, No. 101. Washington, DC: Government Printing Office.

JUDKINS, D., BERK, M., EDWARDS, S., MOHR, P., STEWART, K., and WAKSBERG, J. (1995). National Health Care Survey: List Verses Network Sampling. Unpublished report. National Center for Health Statistics.

JUDKINS, D., MARKER, D., WAKSBERG, J., BOTMAN, S., and MASSEY, J. (1999). National Health Interview Survey: research for the 1995-2004 redesign. National Center for Health Statistics. *Vital and Health Statistics*, Series 2, No. 126, 76-80. Washington, DC: Government Printing Office.

LEAVER, S., and VALLIANT, R. (1995). Statistical problems in estimating the U.S. consumer price index. In *Business Survey Methods*, (B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott Eds.). New York: John Wiley and Sons, Inc.

MASSEY, J.T., MOORE, T.F., PARSONS, V., and TADRO, W. (1991). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics. *Vital and Health Statistics*, Series 2, No. 110. Washington, DC: Government Printing Office.

MATHIOWETZ, N. (1987). Linking the National Survey of Family Growth with the National Health Interview Survey: analysis of field trials. National Center for Health Statistics. *Vital and Health Statistics*, Series 2, No. 103. Washington, DC: Government Printing Office.

SHIMIZU, I., and SIRKEN, M. (1998). More on population based establishment surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 7-12.

SIRKEN, M., SHIMIZU, I., and JUDKINS, D. (1995). The population based establishment surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1, 470-473.

SIRKEN, M. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

THOMPSON, S. (1992). *Sampling*. New York: John Wiley and Sons, Inc.

WUNDERLICH, G.S. (Ed.) (1992). *Toward a National Health Care Survey: A Data System for the 21st Century*. Washington, DC: National Academy Press.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 193-203
Statistics Canada

Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques

JEAN-CLAUDE DEVILLE[1]

ABSTRACT

In a sample survey, in the absence of external information, the total of a variable is estimated using the Horvitz-Thompson estimator. Its variance is in turn estimated by calculating a fairly complex quadratic form, generally recursively. In this paper, this problem is assumed to be solved on the basis of a software capable of carrying out the calculation automatically. In the case of complex estimators (*i.e.*, of the calibration type), and in that of non-linear statistics (substitution estimators), it is shown that the same tool may always be used provided an appropriate artificial variable is chosen. In all cases, this artificial variable provides an estimation of the variance that is approximately unbiased and constructed using the influence function technique as well as some asymptotic postulates. Many examples are provided for the use of this technique: complex but explicit functions of totals (correlation coefficient), implicit functions of totals, calibrated estimators, fractiles and rank statistics, statistics derived from factorial methods.

KEY WORDS: Variance estimation; Complex statistics; Linearization; Substitution estimators; Residual technique; Influence function; Implicit parameters; Fractiles; Rank statistics; Factorial analysis.

1. INTRODUCTION

The formulation of results in the form of confidence intervals is the goal (rarely reached) of all sample surveys. The most common procedure consists in estimating the variance of the statistics involved for the probability distribution induced by the sampling scheme (and sometimes, for the sake of simplicity, following fairly drastic assumptions called models). Then, following the assumption, rarely contradicted by the facts when the samples are large enough, that the statistic follows a normal distribution, a confidence interval symmetric about the point estimation is derived according to simple, standard procedures.

There is abundant literature dealing with this problem, before and after the benchmark work found in the book by Wolter (1985).

The goal of this paper is to show how simple tools can be used to effectively carry out a variance estimation in complex cases by means of a unique technique, *i.e.*, linearization. We will first describe the state of the art concerning the estimation of the variance of the Horvitz-Thompson estimator for a total. After providing a definition of "linearizable statistic" and a description of the concept of influence function analogous to that used in non-parametric statistics, we will introduce the class of functional substitution estimators shown to be linearizable under fairly general assumptions. We will show how the usual rules of differential calculus can be extended to linearized variables, and how, using step-by-step procedures, they can be used to calculate fairly easily the linearized variables of fairly complex statistics. Special attention will be given to statistics using quantiles as well as those linked to the most current multivariate analysis.

This procedure is the chief component of the POULPE software used at INSEE:

- Having a tool to estimate the variance of a total using a simple expansion estimator.
- Reverting to this case, using specially constructed variables, when using a complex estimator and/or when estimating a complex statistic.

2. GENERAL FRAMEWORK: SIMPLE EXPANSION ESTIMATOR

Let us consider a population U of units $k, l, \dots,$ for which a sample design is defined, *i.e.*, a probability distribution p that associates with any part s of U – the sample – a probability $p(s)$ of being selected. Using the latter, probabilities of inclusion π_k $(\pi_k = \sum_{s \ni k} p(s))$ are defined, as are probabilities of inclusion of order $2, \pi_{kl}$ for elements of s. It is then possible to use the Horvitz-Thompson estimator, $\hat{Y} = \sum_{k \in s} y_k / \pi_k$ of the total Y of a variable of interest y. It offers the advantage of being (almost) always available and unbiased, and its variance is easily calculated:

$$\operatorname{Var}(\hat{Y}) = \sum_U \pi_k (1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + 2 \sum \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \quad (2.1)$$

where the second sum extends to all pairs (k, l) of population U.
A useful estimator of the variance of $\hat{Y}$ is given by:

$$\widehat{\operatorname{Var}}(\hat{Y}) = \sum_s (1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + 2 \sum \sum_s \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}} \right) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}. \quad (2.2)$$

[1] Jean-Claude Deville, Laboratoire de Statistique d'Enquête, Insee/Ensai/Crest, Campus de Ker-Lann, 35170-Bruz, France.

In practice, for large samples, this variance estimator is calculated recursively since the double sum which appears has a prohibitive number of terms. Moreover, the probabilities of inclusion π_{kl} can be calculated easily only in some rare simple cases.

In fact, all known sample designs boil down to a few simple schemes: Bernoulli or Poisson sampling, simple random sampling, systematic sampling and sampling with unequal probabilities of fixed size. For the first of these, there are some closed formulas providing a variance estimate based on the sum of squares. The same applies to systematic sampling given a few assumptions that are easily verified for selection order. Finally, the variance of sampling with unequal probabilities of fixed size, for many selection methods, can be approximated in an extremely fine and general manner using the following formula, applied in POULPE (Deville 1993):

$$\hat{V} = \frac{1}{1 - \sum_s a_k^2} \sum_s (1 - \pi_k) \left(\frac{y_k}{\pi_k} - A \right)^2 \tag{2.3}$$

where $a_k = \dfrac{1 - \pi_k}{\sum_s (1 - \pi_k)}$ and $A = \sum_s a_k \dfrac{y_k}{\pi_k}$.

These simple schemes can be combined to provide arbitrarily complex designs by means of two operations, *i.e.*, stratification and multi-stage sampling (or subsampling).

In terms of stratification, a variance estimate of the grand total can be obtained by adding the variances of the estimators of stratum totals.

Multi-stage sampling can be obtained when the population is divided into sub-populations U_i called "primary units". A sample s_1 of the latter is selected on the basis of a sample design p_1 applied to the population of primary units. Then, for each $U_i (i \in s_1)$, a sample is selected using a design p_i / s_1. Conditionally on s_1, these designs are independent. From them are derived the probabilities of inclusion and the following variance formula:

$$\text{Var}(\hat{Y}) = \text{Var}\left(\sum_{s_1} \frac{Y_i}{\pi_i} \right) + E\left(\sum_{s_1} \frac{\text{Var}(\hat{Y}_i)}{\pi_i^2} \Big/ s_1 \right) \tag{2.4}$$

where Y_i is the total of y for U_i, π_i its probability of inclusion, $\hat{Y}_i$ the estimator of Y_i for design p_{i/s_1}. Finally, if $V(Y_i; i \in s_1)$ is the variance estimator of $\sum_{s_1} Y_i / \pi_i$ and V_i a variance estimator of $\hat{Y}_i$ conditionally on s_1, then

$$\widehat{\text{Var}}(\hat{Y}) = V(\hat{Y}_i; i \in s_1) + \sum_{i \in s_1} \frac{V_i}{\pi_i} \tag{2.5}$$

is a variance estimator of $\hat{Y}$ (Durbin 1953).

Naturally, for each stratum h or each primary unit i, the sample survey p_h or p_i / s_1 can itself be stratified or become a multi-stage sampling. However, in all cases, the repetitive and recursive use of the abovementioned rules makes it possible to calculate a variance estimator using simple elements based on the sum of squares. For surveys carried out among people, it is customary for a sample design to comprise three to five selection stages.

This means that the quadratic form (2.2) can be calculated mechanically, without however any explicit computation of the terms involved in the double sum found in the formula.

To complete this overview, it should be noted that a sample is frequently selected in several stages, normally two or three. This means that a sample s is selected and used as a reference population for the selection of a second sample r, using a sample design $q(r/s)$. If it is controlled by the statistician, this design is generally a stratified design with simple random sampling for each stratum. Otherwise, the design is described using a response model that makes it possible to formalize a reweighting procedure for non-response. In all cases, there are second-stage probabilities of inclusion P_k and P_{kl} describing the inclusion in r of the unit k or of the pair (k, l). The expansion estimator is $\hat{Y}_{\text{exp}} = \sum_r y_k / \pi_k P_k$.

Its variance can be calculated fairly easily, and is estimated using the expression:

$$\widehat{\text{Var}}(\hat{Y}_{\text{exp}}) = \sum \sum_r \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}} \right) \frac{1}{P_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum \sum_r \left(1 - \frac{P_k P_l}{P_{kl}} \right) \frac{y_k}{P_k \pi_k} \frac{y_l}{P_l \pi_l} \tag{2.6}$$

with $\pi_{kk} = \pi_k$ and $P_{kk} = P_k$.

In spite of a few difficulties, this variance estimator can be calculated mechanically while avoiding the prohibitive double sums. The same applies to three-stage designs which occur when a non-response stage is added to a second stage controlled by the statisticians. Such procedures are used in POULPE.

Thus, the Horvitz-Thompson estimator (or its extension, the expansion estimator) has a variance that takes on the quadratic form $Q(y_k; k \in U)$. The latter can be estimated without bias (or eventually with negligible bias) using the recursive calculation of a quadratic form $\hat{Q}(y_k; k \in s)$ (where s now represents the final sample, no matter how many stages were needed to obtain it). In the following, we will assume the availability of an "automatic" method of calculating this quadratic form.

3. COMPLEX STATISTICS AND ASYMPTOTIC POSTULATES

We will show that it can also be applied when we use a more refined estimator than HT (involving external auxiliary information), *e.g.*, for complex statistics (means, quantile ratios, complex indexes such as GINI, the coefficient of an econometric model, a principal component analysis factor).

The results we will provide are "asymptotic" approximations. As in Isaki and Fuller (1982) or in Deville and Särndal (1992), we postulate the following scheme: in a series of sampling problems indexed by an integer v (which we will suppress so as not to overload the notation), the size N of the population tends towards infinity as does the size n (or the size expectation) of the sample. For each v , we thus also have a sample design, the associated HT estimator, a vector of invariable fixed size for variables x_k of X estimated using $\hat{X}$.

The three following propositions are postulated:

- $N^{-1}X$ has a limit. (3.1)
- $N^{-1}(\hat{X} - X)$ converges in terms of probability towards zero. (3.2)
- $n^{-1/2}N^{-1}(\hat{X} - X)$ has as a limit a multidimensional normal distribution. (3.3)

The first postulate formalizes the concept of a series of populations of increasing size extracted from a parent continuous distribution. This can also be interpreted as if the population were an i.i.d. sample of a certain infinite superpopulation. The other two postulates relate to the convergence of the HT estimator and to the fact that it leads to a central limit theorem. In practical terms, these postulates are satisfied in many cases, given certain technical assumptions: simple random sampling (Hájek 1964), Poisson sampling and randomized systematic sampling - (Rosen 1972), stratified design with the number of strata tending towards infinity (immediate application of Lindeberg conditions).

In reality, however, we can never tell whether, for example, the design used involves a number of strata tending towards infinity! What the asymptotic postulates mean is simply that certain magnitudes (technically those which are $O_p(n^{-1/2})$) are considered "small" and that the product of the two "small" quantities is a "negligible" (and therefore neglected!) quantity.

On the basis of these postulates, we will show how certain estimators and certain non-linear statistics can be approximated using HT statistics having the form $\sum_s z_k/\pi_k$ for well-chosen variables z_k.

4. SUBSTITUTION ESTIMATORS AND FUNCTIONALS

Let us now consider a fairly general class of non-linear statistics of the finite population based on the concept of a measurement functional, as well as their substitution estimators.

With each unit k of the population U there is associated a point x_k of $\boldsymbol{R}^p$ for the p problem variables of interest to us. The population U is thus represented by the measure M having a unit mass in each of the points x_k.

This measure is positive, discrete and finite, and its total mass has a value of N. We assume that all the x_k are separate, without loss of generality (we can always add a dimension which is the "rank" of k in arbitrary numbering). For any variable $y_k = y(x_k)$, we thus have $\int y dM = \sum_U y_k$.

From an asymptotic point of view, the series of populations is a series of measures on $\boldsymbol{R}^p$. According to the first asymptotic postulate, this series behaves as if we were dealing with i.i.d. selections for a fixed probability distribution on $\boldsymbol{R}^p$.

A functional $T(M)$ associates with any measure of a class containing at least the point measurements, a real number or a vector. We also assume that all the functionals of interest are homogeneous, *i.e.*, there is a positive real number α dependent on T such that $T(tM) = t^\alpha T(M)$ for any positive real number t. A total is a homogeneous functional of level 1, a mean of level 0, a sum having a double index of level 2. Being limited to homogeneous functionals is not too cumbersome in practical terms.

Now let $\hat{M}$ (estimator of M) denote the measure allocating a weight w_k to any point x_k for k in s and zero to any other point, regardless of the origin of the weights (Horvitz-Thompson or calibration).

Definition: The substitution estimator of a functional $T(M)$ is $T(\hat{M})$.

In the case of a total, this definition should not be surprising since $T(\hat{M}) = \int x d\hat{M}(x) = \sum_s x_k w_k$. For "ordinary" complex statistics (ratios, means or indexes, for example), this represents the common practices of survey operations. The same applies to statistics of rank, with finer points having more to do with the estimation of the distribution function than the estimation of the fractiles (see for example Chambers, Dorfmann and Hall 1992).

A fairly general class of parameters linked to the finite population can be obtained using implicit equations which define them. Such is the case, for example, for the adjustment of a parametric model at the population level leading to an "estimating equation" derived from a broad adjustment principle (maximizing likelihood, minimizing chi-2 or "moment" methods or "generalized moment" methods).

This form of writing introduces the (eventually multidimensional) model parameter as a functional of M. Its estimator (in the sense of sampling) is the same functional for $\hat{M}$. Thus, the estimation of the least squares in the linear model is written as follows:

$$B = \arg\ \operatorname{Min}\sum_U q_k(y_k - x_k'B)^2.$$

The estimation (sampling) of B is $\hat{B} = \arg\ \operatorname{Min}\sum_s w_k q_k (y_k - x_k'B)^2$. The use of $\hat{B}$ (rather than an estimator for a model conditional upon the sample) is much more robust, and correctly accounts for the fluctuations of sampling on the result (on this point, see Binder 1983, Binder and Patak 1994, or Binder and Kovačević 1997).

Generally speaking, an "estimating equation" at the population level will be written as $T(M, \lambda) = 0$ where T is a functional of dimension p parametered by vector λ also of dimension p. This equation will be assumed to have a unique solution for fixed M. The substitution estimator of λ is the solution of the (estimating) equation $T(\hat{M}, \hat{\lambda}) = 0$.

5. LINEARIZABLE STATISTIC

Let us consider some statistics S dependent on the observations $(x_k; k \in s)$ (in fact a series of statistics defined in each of the sampling problems within the asymptotic framework). S is said to have a probability of order $f(n)$ (where f is some positive function of n), and we write $S = O_p(f(n))$ if for any $\varepsilon > 0$ there is a constant C such that

$$\Pr\left(\frac{\|S\|}{f(n)} \geq C\right) \leq \varepsilon.$$

In other words, the survivorship functions of variables $\|S\|/f(n)$ are uniformly overestimated by the survivorship function represented as $(C(\varepsilon), \varepsilon)$.

The third asymptotic axiom (central limit axiom!) can therefore be written as $N^{-1}(\hat{X} - X) = O_p(n^{-1/2})$, and the second as $N^{-1}\hat{X} = O_p(1)$. In the rest of this paper, we will use more or less implicitly the following well-known result (see for example Billingsley 1969):

Result: If a statistic S converges towards a certain distribution, and if $(S - T) = O_p(f(n))$ with $f(n) \to 0$, then T converges towards the same distribution. Specifically, S and T have the same limit variance.

The statistic S is said to be of degree α if $N^{-\alpha}S$ tends towards a limit. Clearly, for example, a HT estimator is of degree 1, a ratio of HT estimators is of degree 0 (or homogeneous). The third asymptotic postulate states that $nE(\hat{X} - X)^2$ is of degree 2. The substitution estimator of a homogeneous functional of degree α is a statistic of degree α.

The following definition can now be formulated:

Definition: A statistic S of degree α is linearizable if there is a synthetic variable z_k (known as the linearized variable of S) such that the variance of $\hat{Z}$ is equivalent to that of S in the sense that $n^{1/2}(N^{-\alpha}S - N^{-1}\hat{Z}) = O_p(f(n))$ with $f(n) \to 0$. In general, we will almost always have $f(n) = n^{-1/2}$.

In practice, this means that the variance, and therefore a confidence interval, will be estimated for S on the basis of the variance of $\hat{Z}$ (whether or not it is a HT estimator).

Note, on the other hand, that the definition does not imply the uniqueness of the variable linearizing a statistic. Specifically, the approximation contained in the definition can be more or less fine at two levels: that of the convergence speed $f(n)$, and, for an equal speed, that of the increment $C(\varepsilon)$.

Generally speaking, however, the linearized variable z_k cannot be computed explicitly by means of data from the sample. We are then led to replace z_k by an approximation $\tilde{z}_k$ using certain statistics estimated on the basis of the sample. This occurs in the most elementary cases. The matter of the legitimacy of this approximation must be dealt with, and this can only be done within an asymptotic framework.

Result: If quantities $\tilde{z}_k$ depend regularly on a fixed, finite number of estimated parameters, then the variance estimators $\hat{Q}(z_k; k \in s)$ and $\hat{Q}(\tilde{z}_k; k \in s)$ are equivalent, *i.e.*, their difference as normalized by factor n/N^2 is an asymptotically negligible quantity.

Proof: By "regularly" is meant that $\tilde{z}_k = z_k + \xi_k'(\hat{\Gamma} - \Gamma) + O_p(\|\hat{\Gamma} - \Gamma\|^2)$, where Γ is the p vector of the parameters, $\hat{\Gamma}$ is its vector of estimators and ξ_k is a p-variable. The asymptotic postulates tell us that $n/N^2 Q(z_k)$ converge towards a finite quantity just as $n/N^2 Q(z_k, \xi_k) = \sum_{k,l} \Delta_{kl}\, z_k\, \xi_l$ if the quadratic form is made explicit, and that $n/N^2 Q(\xi_k)$. We then have:

$$Q(\tilde{z}_k) = Q(z_k) + 2Q(z_k, \xi_k)'(\hat{\Gamma} - \Gamma) + O_p(\|\hat{\Gamma} - \Gamma\|^2).$$

As $\|\hat{\Gamma} - \Gamma\| = O_p(n^{-1/2})$, we obtain the result.

When the number of estimated parameters tends towards infinity, the situation is not perfectly clear. In practical terms, obviously, what is meant by the number of estimated parameters tending towards infinity? Theoretically, moreover, there are some difficulties as can be seen from the following two contradictory examples:

Example: Poststratification. We assume the poststrata defined on the basis of a numerical variable, and we construct m adjacent poststrata, each of which comprises about n/m surveyed units. Here vector Γ is that of the m means of poststrata $\bar{Y}_h$, $h = 1$ at m. If m increases with n, each estimated parameter $\hat{\bar{Y}}_h$ is such that $\hat{\bar{Y}}_h - \bar{Y} = O_p((n/m)^{-1/2})$. Then $\|\hat{\Gamma} - \Gamma\|$ is of the order of $m^{3/2} n^{-1/2}$. Taking $m = n^{\alpha}$ with $\alpha < 1/3$, the previous result and its proof remain valid.

Counter-example: For the estimation of inequality indexes, we are led to use statistics such as $S = \sum_s y_k \hat{F}(y_k)$ where $\hat{F}$ is an estimation of the distribution function of variable y. If R_k denotes the rank of y_k in the population, we could imagine that $z_k = 1/N\, y_k R_k$ is a linearized statistic for S. This is completely false (Deville 1997).

The difference with respect to the previous example rests in the fact that S uses an estimated parameter per sampled unit, in which case anything can happen! The general procedure for dealing with such statistics will be described below in section 12.

6. INFLUENCE FUNCTION OF A FUNCTIONAL AND ASYMPTOTIC VARIANCE OF THE SUBSTITUTION ESTIMATOR

Definition: The influence function (if it exists) of a functional T is:

$$IT(M;x) = \lim_{t \to 0} \frac{1}{t}(T(M + t\delta_x) - T(M))$$

where δ_x denotes the unit mass assumed at point x.

Comment: This definition is slightly different from that used in the field of robust statistics (Hampel, Ronchetti, Rousseeuw, and Stahel 1985). It is made necessary by the fact that the total mass of M is variable, and often unknown in a statistical problem. It is nothing more than the differential as viewed by Gateaux for a Dirac mass assumed at a point x.

The essential point of this paper can now be formulated:

Result: Under broad assumptions, the substitution estimation of a functional $T(M)$ is linearizable. A linearized variable is $z_k = IT(M;x_k)$ where IT is the influence function of T in M.

Comment: The influence function can thus be used to estimate the variance of $T(\hat{M})$. This being said, very often the influence function includes in its definition certain functionals of M (*e.g.*, a ratio or a mean). We are thus led to choose an estimation of the influence function itself in order to compute the variance estimation. This choice is not necessarily unique.

Proof of the result: Let us provide the space of measurements on $\boldsymbol{R}^q$ with a metric d accounting for the convergence: $d(M_1, M_2) \to 0$ if and only if $N^{-1}(\int y dM_1 - \int y dM_2) \to 0$ for any variable of interest y. The asymptotic postulates mean that $d(\hat{M}/N, M/N)$ tends towards zero. We can visibly ensure that $d(\hat{M}/N, M/N)$ is $O_p(1/\sqrt{n})$ according to the third postulate. Now, let us assume that T can be derived in accordance with Fréchet, *i.e.*, for any direction of the increase, in the space of "useful" measures provided with the abovementioned metric. Thus we have:

$$N^{-\alpha}(T(\hat{M}) - T(M)) = \frac{1}{N}\sum_U z_k(w_k - 1) + o\left(d\left(\frac{\hat{M}}{M}, \frac{M}{N}\right)\right).$$

The result is that:

$$\sqrt{n}N^{-\alpha}(T(\hat{M} - T(M))) = \frac{\sqrt{n}}{N}\sum_U z_k(w_k - 1) + o_p(1).$$

Thus, according to the third postulate, the variance of the second member tends towards a limit, that of $n/N^2 \operatorname{Var}(\hat{Z})$, and the result is obtained.

7. EXAMPLES AND COMPUTING RULES FOR INFLUENCE FUNCTIONS

Example 1: If T is the total $T = \int x dM(x)$ of a variable, the influence function of T is this variable itself: $IT(M, x) = x$. Specifically, if $x = 1$, $T = N$ the population size. The influence function is then constant, and its value is 1.

The rules of composition for influence functions are copied from those of differential calculus:

Rule 1: If f is a derivable function defined on the space of values for T a vector function, we have:

$$I(f(T)) = Df(T)\, IT$$

(where Df represents the matrixes of the partial derivatives of f).

The proof is immediate.

Example 2: $f(T) = 1/T$ and $T = \int x dM$, scalar total. The influence function is $-x/T^2$.

Rule 2: If S and T are two functionals, we have:

$$I(S + T) = IS + IT \quad \text{and} \quad I(ST) = SIT + TIS.$$

If T and S have vector values, and if H is a matrix, we have, when the products are defined:

$$I(HT) = HIT \quad \text{and} \quad I(S'HT) = (IS)'HT + S'HIT.$$

Example 3: $R = \int y dM \big/ \int x dM = Y/X$ a ratio of two totals. The influence function is:

$$\frac{y}{X} - \frac{Yx}{X^2} = \frac{1}{X}(y - Rx).$$

For a mean $\bar{Y} = \int y dM / \int dM$, the influence function is therefore: $1/N(y - \bar{Y})$, which is the usual definition, or just about, given in the robustness theory (Lecoûtre and Tassi 1987).

Rule 3: Let $S_i (i = 1, ..., q)$ denote scalar functionals and $S = \Pi_{i=1}^q S_i$. We have:

$$IS = S\left(\sum_{i=1}^{q} \frac{IS_i}{S_i}\right).$$

Proof: $I(\mathrm{Log}S) = \dfrac{IS}{S}$.

Now let $T(\lambda) = T(M, \lambda)$ denote a family of functionals depending regularly on a parameter λ that varies in a domain of $\boldsymbol{R}^q$, with Λ a measure on this domain. This leads to:

Rule 4: $I(\int T(\lambda) d\Lambda(\lambda)) = \int IT(\lambda) d\Lambda(\lambda)$. This is elementary.

Note: The persistency conditions include the possibility of reaching the limit under the integration sign.

If, moreover, φ is a function of $\boldsymbol{R}^q$ in the domain of T measurable for all measures M of interest, we proceed as follows:

Rule 5: $I(\int T(\varphi(x))dM(x);\xi) = T(\varphi(\xi)) + \int IT(M,\varphi(x);\xi)$ $dM(x)$.

Proof: (provided as an example): Let $S(M) = \int T(M, \varphi(x))dM(x)$. We have:

$$\frac{1}{t}\left[s(M + t\delta_\xi) - S(M)\right] = \frac{1}{t}\int\left[T(M + t\delta_\xi, \varphi(x)) - T(M, \varphi(x))\right]$$
$$dM(x) + \delta_\xi(T)(T(M + t\delta_\xi), \varphi(x)).$$

The second term tends towards $T(M, \varphi(\xi))$ whenever t tends towards zero. The former can be written as follows:

$$\left[\int IT(M,\varphi(x);\xi) + R_{M,\varphi(x);\xi}(t)\right]dM(x)$$

where R is a quantity that tends towards zero (it may be assumed that the convergence is consistent at x). The result is derived immediately.

Let us now assume that $T(\lambda)$ is a functional with values in $\boldsymbol{R}^q$, regular at λ. Specifically, then, matrix $\partial T / \partial\lambda$ is reversible for any M, and, for fixed M, application $\lambda \rightarrow T(\lambda)$ is one-one and allows for a partial reciprocal function. Equation $T(\lambda) = T_0$ therefore has a unique solution for any M, defining a functional $\lambda(M)$.

Rule 6: The influence function of $\lambda(M)$ is:

$$I\lambda(M;\xi) = -\frac{\partial T}{\partial\lambda}(M,\lambda)^{-1}\, IT(M,\lambda;\xi).$$

Proof: $T(M + t\delta_\xi, \lambda(M + t\delta_\xi) - T(M,\lambda)) = 0$, hence:

$$IT(M,\lambda;\xi) + \frac{\partial T}{\partial\lambda} I\lambda(M;\xi) = 0.$$

This rule may also be needed:

Rule 7: Let S denote a functional in $\boldsymbol{R}^q$, and let T_λ denote a family of functionals regularly indexed by $\lambda \in \boldsymbol{R}^q$. We have:

$$I(T_s) = IT_{\lambda/\lambda = s} + \left(\frac{\partial T}{\partial\lambda}\right)_{\lambda = s} IS.$$

Proof: Writing everything, we have $I(T_s) = IT(S(M), M)$. The rest is obvious.

Note, finally, the interesting link between the influence function and the functional from which it is derived:

Result: If T is homogeneous of degree α we have:

$$\int IT(M;x)dM(x) = \sum_U IT(M;x_k) = \alpha T(M).$$

The specific case $\alpha = 0$ shows that any homogeneous functional has a zero-sum influence.

Proof: We have:

$$\frac{T((1 + h)M) - T(M)}{h} = \frac{((1 + h)^\alpha - 1)}{h} T(M)$$

from the definition of homogeneity. The result follows from the linearity of the derivation as interpreted by Gateaux and the definition of influence function.

8. APPLICATIONS: FUNCTIONS OF TOTALS

We have already seen that, for linear functionals, $T(M) = \sum_U y_k = \int y dM$, the influence function is y_k itself. The application of the notion of influence function becomes redundant. It will be noted, however, that it is in no way asymptotic.

Function of totals: If X is a vector of totals, the influence function of X is, naturally, the vector x_k of the variables making up X. As a result, if $T(M) = f(X) = f(\int x dM)$, the influence function of T is:

$$IT(M;x_k) = f'(X).\, IT(\int x dM) = f'(X).x_k$$

where $f'(X)$ is the row vector of the partial derivatives of f with respect to the coordinates of X taken at point X. We are led naturally to the classical result of Woodruff (1971).

In line with the above, the substitution estimator of $f(X)$ is $f(\hat{X})$. Its approximate variance is that of $f'(X).x_k$, and it is numerically approximated by $f'(\hat{X}).x_k$ in compliance with common practices.

Example: The ratio $R = f(X, Y) = Y/X$ of two scalar totals is estimated using $\hat{R} = \hat{Y}/\hat{X}$. This statistic (of degree 0) allows as a linearized variable $z_k = 1/X(y_k - Rx_k)$. To numerically compute the variance estimation of $\hat{R}$, we use the approximation $\tilde{z}_k = 1/\hat{X}(y_k - \hat{R}x_k)$, an expression which depends on $\hat{Y}$ and $\hat{X}$ and therefore on s; $\tilde{z}_k$ is therefore not a linearized variable as understood in the definition.

Example: Ratio estimator.

It is $\hat{Y}_{\text{rat}} = (X/\hat{X})\hat{Y}$. If we refer to the previous example, the linearized variable of $\hat{Y}_{\text{rat}}$ is $y_k - Rx_k$, approximated by $y_k - \hat{R}x_k$. And yet it could also be said that the estimated variance of $\hat{Y}_{\text{rat}}$ must be equal to X^2 times the estimated variance of $\hat{R}$, which leads to the approximation $X/\hat{X}(y_k - \hat{R}x_k)$ which has many times been deemed more interesting than the previous one. This example shows that the choice of linearized variable is not necessarily unique once external information is used.

Nevertheless, one of the advantages of the influence function approach is to provide computations fairly easily in apparently complex cases.

Example: The correlation coefficient between x and y is written as follows:

$$\rho = \frac{\sum_U x_k y_k - \frac{1}{N}\sum_U x_k \sum_U y_k}{\sqrt{\sum_U x_k^2 - \frac{1}{N}\left(\sum x_k\right)^2}\sqrt{\sum_U y_k^2 - \frac{1}{N}\left(\sum y_k\right)^2}}$$

$$= \frac{V_{XY}}{\sqrt{V_{XX}V_{YY}}}$$

Using the logarithmic derivatives (rule 3), we obtain:

$$\frac{(I\rho)_k}{\rho} = \frac{I(V_{XY})_k}{V_{XY}} - \frac{1}{2}\frac{I(V_{XX})_k}{V_{XX}} - \frac{1}{2}\frac{I(V_{YY})_k}{V_{YY}}.$$

The influence of $A_{XY} = \frac{1}{N}\sum_U x_k \sum_U y_k$ is obtained in the same way using:

$$I(A_{XY})_k = A_{XY}\left(\frac{x_k}{X} + \frac{y_k}{Y} - \frac{1}{N}\right) = \bar{Y}x_k + \bar{X}y_k - \bar{X}\bar{Y}$$

hence: $I(V_{XY})_k = x_k y_k - I(A_{XY})_k = (x_k - \bar{X})(y_k - \bar{Y})$.
Then $I(V_{XX})_k = (x_k - \bar{X})^2$, and $I(V_{YY})_k = (y_k - \bar{Y})^2$ and so, with

$$S_x^2 = \frac{V_{XX}}{N} \text{ and } \text{«}_Y^2 = \frac{V_{YY}}{N}:$$

$$N(I\rho)_k = \frac{(x_k - \bar{X})(y_k - \bar{Y})}{S_X S_Y} - \frac{1}{2}\rho\left(\frac{(x_k - \bar{X})^2}{S_X^2} + \frac{(y_k - \bar{Y})^2}{S_Y^2}\right).$$

And the work is done.

9. APPLICATION: IMPLICIT PARAMETER

Let us assume that B, a parameter with q components, is a solution to an equation of the type:

$$H(B) = \sum_U I_k(B) = 0 \qquad (9.1)$$

where the I_k are regular functions of $\boldsymbol{R}^q$ in $\boldsymbol{R}^q$. This situation frequently occurs when B is the parameter of a model assumed to be valid in the population U. Under the usual assumptions of independence, equation (9.1) can result from the application of the maximum likelihood estimation principle. It is then the equation of the score. In the case of a linear model with Gauss residuals, this leads to the normal equations that can also be derived from the least squares principle:

$$I_k(B) = \frac{1}{\sigma_k^2}x_k(y_k - x_k'B)$$

with obvious notations.
If the functional family $H(B)$ regularly depends on B, we have:

$$I(B_0)_k = -\left(\frac{\partial H}{\partial B}\right)^{-1}_{B = B_0} IT(B_0)_k$$

i.e.,:

$$I(B_0)_k = -\left(\sum_U \frac{\partial I_k}{\partial B}\right)^{-1} I_k(B_0).$$

In the case of regression, we have:

$$I(B)_k = -\left(\sum_U \frac{x_k x_k'}{\sigma_k^2}\right)^{-1}\frac{1}{\sigma_k^2}x_k e_k = -T^{-1}\frac{1}{\sigma_k^2}x_k e_k$$

with the regression residual e_k. Thus we simply have the linearized variable of the vector of regression coefficients. To numerically compute the variance estimator, we use the approximation

$$\tilde{z}_k = \hat{T}_s^{-1}\frac{1}{\sigma_k^2}x_k\tilde{e}_k \text{ where } \hat{T}_s = \sum_S \frac{x_k x_k'}{\sigma_k^2\pi_k}$$

and $\tilde{e}_k = y_k - \hat{B}x_k$. This expression therefore depends on s through $\hat{T}_s$ and

$$\hat{A} = \sum_s \frac{x_k y_k}{\sigma_k^2\pi_k}.$$

Example: Regression estimator.

When the constant (or the variable σ_k^2) is part of the regressors, *i.e.*, when there is a vector λ such that $x_k'\lambda = 1$ (or σ_k^2) for any x, the regression estimator takes on the simple form $\hat{Y}_{\text{reg}} = X'\hat{B}$ where X is the known vector of the total of the x_k.

Regression estimation theory (Cochran 1977, Särndal, Swensson, Wretman 1992) tells us that the residuals e_k are the linearized variable of this estimator, and that they can be approximated using the estimated empirical residuals $\tilde{e}_k$ (note that these only depend on a finite number of parameters).

However, the above leads us to think that we should have:

$$\text{V}\hat{\text{a}}\text{r}(\hat{Y}_{\text{Reg}}) = X'\text{V}\hat{\text{a}}\text{r}(\hat{B})X$$

and that a natural "linearized" variable for $\hat{Y}_{\text{Reg}}$ should be $X'\hat{T}_s^{-1}1/\sigma_k^2 x_k\tilde{e}_k$. If $\hat{T}_s$ is replaced by its expectation T, we notice that $X'T^{-1} = \lambda'$ and that $1'x_k = \sigma_k^2$, and we fall back on the previous approximation. Note, finally, that the quantities $X'\hat{T}_s^{-1}$ are exactly the weight corrections (or g – weights) used in the regression estimator, the use of which is often recommended within the framework of variance estimation.

It is quite clear that the two linearized variables lead asymptotically to the same result. The choice should therefore be based on other criteria. In a few specific cases, the concept of conditional estimation justifies the use of these g – weights, notably for the poststratified estimator in the case of simple random sampling. The general case remains fairly mysterious.

In the case of a logistic regression adjustment, the dependent variable y_k has a value of 0 or 1, and the equations of the score are written as follows:

$$H(B) = \sum_U x_k(y_k - f(x_k'B)) \text{ with } f(u) = \exp u/(1 + \exp u).$$

We therefore have:

$$I(B)_k \equiv \left(\sum_U x_k x_k' f(x_k'B)(1 - f(x_k'B)) \right)^{-1} x_k (y_k - f(x_k'B)).$$

Using this variable makes it possible to compute correctly the precision of a logistic regression, *i.e.*, taking into consideration the sampling scheme.

10. THE RESIDUAL TECHNIQUE FOR COMPLEX ESTIMATORS

Many complex estimators commonly used nowadays can be incorporated into the general framework of external data calibration (Deville, Särndal 1992; Deville, Särndal, Sautory 1993). A vector X of totals of auxiliary variables x_k is known, and we look for new weights w_k confirming the calibration equations $\sum_s w_k x_k = X$ for any sample s. If we look for such weights as close as possible to the HT weights, they are found to be necessarily of the type $w_k = 1/\pi_k F_k(x_k'\lambda)$ where λ is a vector of the same dimension as X solving the calibration equations. The functions F_k depend on the chosen distance and allow limited development in the form $F_k(u) = 1 + q_k u + O(u^2)$. The most frequent form is $F_k(u) = F(q_k u)$, F a unique function. Often, also, the q_k are all equal to 1. Thus we find in this family the ratio estimator (arbitrary F and $q_k = 1/x_k$), the poststratified estimator (with x_k a stratum indicating vector), the raking ratio estimator (with x_k a margin indicating vector and F an exponential function), and the regression estimator ($F(u) = 1 + u$).

The asymptotic variance of these estimators can be obtained naturally by applying the rules of linearization. The calibration equations define λ using:

$$T(\hat{M}, \lambda) = \int x_k F_k(x_k'\lambda) d\hat{M}(k) = \sum_s \frac{1}{\pi_k} x_k F_k(x_k'\lambda) = X$$

Since we have $T(M, 0) = X$, the application of rule 6 yields:

$$I\lambda(M, x) = -T^{-1} x_k$$

with

$$T = \int x_k x_k' F_k(0) dM(k) = \sum_U q_k x_k x_k'.$$

Moreover, the calibrated estimator appears to be the substitution estimator of the functional $S(M, \lambda) = \int y_k F_k(x_k'\lambda)\ dM(k)$, which, according to rule 7, allows for the linearized variable $y_k - x_k' T^{-1} \int q_i y_i x_i dM(i) = y_k - x_k' B$ by introducing the vector B of the least squares regression parameters into the population for the weights q .

Thus the variance of the calibrated estimator is obtained by replacing, in formula (2.1), the y_k by the residuals $e_k = y_k - x_k' B$ of the regression of y on x with the weights q. For the variance estimation, we use in formula (2.2) either $\tilde{e}_k = y_k - x_k' \hat{B}$, or, as in the case of the regression estimator, $F(x_k'\lambda)\tilde{e}_k$.

If we now turn to a parameter $T(M)$ estimated by substitution using the weights w_k obtained by calibration, we have the following important result:

Result: If $T(M)$ allows for a linearized variable z_k, and if $T(\hat{M}_w)$ is the estimator of $T(M)$ using the weights w_k derived from calibration on a vector $X = \sum_U x_k$, then e_k, a residual of the regression z_k on x_k is a linearized variable for $T(\hat{M}_w)$.

Proof: The variance of $T(\hat{M}_w)$ is equivalent to that of $\hat{Z}_w = \sum w_k z_k$ according to the previous demonstration. However, the variance of $\hat{Z}_w$ is equivalent to that of $\sum_s 1/\pi_k e_k$.

Comment: Very often, *e.g.*, in the case of an explicit function of totals, z_k is a linear form $\sum_{i=1}^p A_i y_k^i$. We then have:

$$\mathrm{Var} \sum_s w_k z_k = \mathrm{Var} \sum_{i=1}^p A_i \hat{Y}_w^i$$

This suggests the following procedure:

- compute the residuals ε_k^i of the regressions of y_k^i on the x_k.
- form the synthetic variable $\sum_i A_i \varepsilon_k^i$
- compute the variance of this variable.

It is quite clear that this corresponds to the direct computation of the residuals of z_k, which is definitely more simple.

Comment: While it may be trivial, this result is perhaps the most useful one in this paper, and this comment simply ensures that it will not go unnoticed.

11. APPLICATION: FRACTILES

The distribution function $F(x) = 1/N \mathrm{Card}\,(k; x_k \le x)$ is a functional family $1/N \int 1(\xi \le x) dM(\xi)$. The value of influence $IF(x)_k$ is therefore

$$\frac{1}{N}(\mathbf{1}(x_k \le x) - F(x)) \qquad (11.1)$$

For $\alpha \in]0, 1[$, the fractile t_α is defined by $F(t_\alpha) = \alpha$ if we are ruthless, and by t_α: $F(t_\alpha - 0) < \alpha \le F(t_\alpha)$ if we take into consideration the staircase-shape of F. If we are ruthless, the ad hoc linearized variable is therefore:

$$I(t_\alpha)_k = -\left(\frac{\partial F}{\partial x} \bigg| x = t\alpha \right)^{-1} IF(t_\alpha)_k$$

$$= -\frac{1}{F'(t_\alpha)} \cdot \frac{1}{N} \left(\mathbf{1}(x_k \le t_\alpha) - \alpha \right) \qquad (11.2)$$

The problem arises from the fact that $F'(x)$ idealizes a density of the variable at point x which does not exist because of the stairs.

The difficulty can be overcome by using the following construction:

By definition, a regulating core is a positive function $K(x, t)$, confirming, for any $x, \int K(x, t)dt = 1$, which is regular (*e.g.*, sufficiently derivable). For any x, $K(x, .)$ is a "bell" function about x, *e.g.*, a normalized indicatrix of an interval surrounding x. More generally, the support of $K(x, .)$ will be an interval containing x. We note $G(x, t) = \int^{t} K(x, u)du$ and $\bar{G}(x, t) = 1 - G(x, t)$. $G(x;.)$ is a distribution function. From an asymptotic point of view, the core K depends on the size N of the population; the "band width", *i.e.*, the "mean" width of the support of $K(x,.)$, decreases with N.

We now replace the distribution function by its smoothing $F_K(x) = \int F(t) K(x, t)dt$. For a reasonable choice of K, F_K is strictly increasing wherever its value is not 0 or 1, and very close to F so that all the fractiles $t_{K\alpha}$ are defined univocally and close to t_α no matter how they are defined. Following integration by parts, note that we also have:

$$F_K(x) = \int \bar{G}(x, t)dF(t) = \frac{1}{N}\sum_{U} \bar{G}(x, x_k).$$

We therefore have:

$$I(F_K(x),\xi) = \frac{1}{N}(\bar{G}(x, \xi) - F_K(x)) \qquad (11.3)$$

which is entirely analogous to (11.1).

Since F_K is derivable ($\bar{G}$ being so), we have:

$$It_{K\alpha}(x) = -\frac{1}{F_K'(t_{K\alpha})}\frac{1}{N}(\bar{G}(t_{K\alpha}, x) - \alpha) \qquad (11.4)$$

This formula is entirely analogous to (11. 2) save that $F_K'(t_{K\alpha})$ is perfectly defined. The linearization of $t_{K\alpha}$ does not therefore cause any particular problem, and may be used approximately for the linearization of t_α itself. A combined strategy consists in using the linearized variable

$$z_k = -\frac{1}{F_K'(t_\alpha)}(\mathbf{1}(x_k \le t_\alpha) - \alpha)$$

with

$$K(x, t) = \frac{1}{b - a}\mathbf{1}(a \le t < b)$$

(where $[a, b]$ is an interval containing x, more or less arbitrary). A practically correct linearized variable would be:

$$z_k = -\frac{b - a}{F(b) - F(a)}(\mathbf{1}(x_k \le t_\alpha) - \alpha).$$

The interval [a, b] will have to be large enough so that in

$$\tilde{z}_k = -\frac{b - a}{\hat{F}(b) - \hat{F}(a)}(\mathbf{1}(x_k \le \hat{t}_\alpha) - \alpha)$$

the first factor will be sufficiently insensitive to sampling fluctuations.

12. INDEXES OF CONCENTRATION AND OTHER FUNCTIONALS LINKED TO RANKS

Let us consider a few examples.

(a) GINI index.

With $T_x = \int \mathbf{1}(\xi < x)dM(\xi)$, the Gini index can be defined as:

$$\text{GINI} = \frac{\int x\, T_x\, dM(x)}{N X}$$

Applying rule 5, we find for the influence of the numerator $xT_x + \int \xi \mathbf{1}(x \le \xi)dM(\xi)$. And yet $\int \xi \mathbf{1}(x \le \xi)dM(\xi) = X - \int \xi \mathbf{1}(\xi \le x)dM(\xi) = X - T_x \bar{x}_<$ where $\bar{x}_<$ is the mean of the x_k lower than x. Since X is a constant, the numerator linearized variable is therefore $T_x(x - \bar{x}_<)$. A linearized variable for GINI is therefore:

$$\text{IGINI}_k = F(x_k)\frac{x_k - \bar{x}_<}{X} - \text{GINI}\frac{x_k}{X}.$$

(b) Population below the poverty threshold.

It is defined as the proportion (of revenues) lower than half the distribution median. For the proper weight, let α and β denote two numbers between 0 and 1, and let us consider the indicator $J_{\alpha\beta} = F(\beta t_\alpha)$. The usual indicator corresponds to $\alpha = \beta = 1/2$.

The linearization is obvious using the rules under section 6 and the convention for writing the distribution function derivative:

$$\begin{aligned} IJ_{\alpha\beta}(x) &= IF_{\beta q_\alpha}(x) + F'(\beta t_\alpha)\beta I_{t_\alpha}(x) \\ &= \frac{1}{N}(\mathbf{1}(x \le \beta t_\alpha) - F(\beta t_\alpha)) - \frac{1}{N}\frac{F'(\beta t_\alpha)}{F'(t_\alpha)}(1(x \le t_\alpha) - \alpha) \\ &= \frac{1}{N}\left[\mathbf{1}(x \le \beta t_\alpha) - \frac{F'(\beta t_\alpha)}{F'(t_\alpha)}1(x \le t_\alpha) + (\alpha - F(\beta t_\alpha))\right] \end{aligned}$$

For $\beta = 1$ we are able to find $IJ_{\alpha 1} = 0$.

The variance of the indicator is therefore computed simply by using the artificial variable having a value of 1 if $x_k \le \beta\hat{t}_\alpha$,

$$1 - \frac{F'(\beta\hat{t}_\alpha)}{F'(\hat{t}_\alpha)}$$

if $\beta\hat{t}_\alpha < x_k \le \hat{t}_\alpha$ and 0 if $x_k > \hat{t}_\alpha$.

(c) Kendall's coefficient of rank correlation.

Two numerical variables x_k and y_k are linked to individual k. The ranks of x_k and y_k respectively can be written as $R_k^X = \int_{x \le x_k} dM(\dot{x}, y)$ and $R_k^Y = \int_{y \le y_k} dM(x, y)$. The coefficient of rank correlation is the correlation coefficient between R_k^X and R_k^Y, *i.e.*, following some elementary simplifications:

$$r = 12\left(\frac{1}{N^3}\int R_\xi^X R_\eta^Y dM(\xi, \eta) - \frac{1}{2}\right).$$

This expression can be linearized by applying the rules related to influence functions. For: $T = \int R_\xi^X R_\eta^Y dM \;\; (\xi, n)$, we have:

$$IT(x, y) = R_x^X R_y^Y + \int \mathbf{1}(x \le \xi) R_\eta^Y dM(\xi, \eta)$$

$$+ R_\xi^X \mathbf{1}(y \le \eta) dM(\xi, \eta)$$

$$= R_x^X R_y^Y + A_x + B_y$$

where we have assumed

$$A_x = \sum_{k \in U : x \le x_k} R_K^Y$$

and

$$B_y = \sum_{k \in sU : y \le x_k} R_k^X,$$

so that finally:

$$Ir(x, y) = \frac{12}{N}\left(F_x^X F_y^Y + \frac{A_x}{N^2} + \frac{B_y}{N^2} - \frac{1}{4}\left(r + \frac{1}{2} \right) \right)$$

The variance is computed as follows:

- The linearized variable is $z_k = Ir(x_k, y_k)$.
- $\hat{F}_{x_k}^X$ and $\hat{F}_{y_k}^Y$ are the estimators of the distribution functions of x and y respectively.
- A_x is estimated using

$$\hat{A}_x = \sum_{k \in s : x_k \ge x} w_k$$

  and B_y likewise.
- In the calculation, we use the approximation of

$$z_k, \tilde{z}_k = \frac{12}{\hat{N}}\left(\hat{F}_{x_k}^X \hat{F}_{y_k}^Y + \frac{A_{x_k}}{\hat{N}_2} + \frac{B_{y_k}}{\hat{N}_2} - \frac{1}{4}\left(\hat{r} + \frac{1}{2} \right) \right)$$

and we calculate the variance of the total of this variable estimated using the HT estimator (formula 2.2).

13. FACTORIAL METHODS

The principal components of the vectorial variable x_k are the eigenvectors u of the matrix of covariances $C = \sum_U x_k x_k' - X\bar{X}'$. They therefore confirm:

$$\begin{matrix} Cu = \lambda u \\ u'u = 1 \end{matrix} \quad \text{with } \lambda \text{ the eigenvalue.}$$

The variance of λ and that of the components of the u can be obtained fairly simply. The influence of C is $IC(x) = (x - \bar{X})(x - \bar{X})'$. The influence of $Cu - \lambda u$ is

$$ICu + CIu \;\; - I\lambda u - \lambda Iu = 0 \qquad (13.1)$$

However $(Iu)'u = 0$, and also $u'CIu \; = 0$ because C is a symmetric matrix. By multiplying (13.1) on the left by u' we have:

$$u'ICu = I\lambda = (u'(x - \bar{X}))^2.$$

And yet $u'(x - \bar{X})$ is equal to $\lambda\xi^2$ where ξ is the principal component associated with (λ, u). From this is derived the calculation of the variance of $\hat{\lambda}$, the solution to $\hat{C}\hat{u} - \hat{\lambda}\hat{u} = 0$.

The variance of the components of u is obtained analogously. Let (λ_ν, ν) denote another eigenvalue, eigenvector pair of C. We multiply equation (13.1) on the left by ν. We have:

$$\nu'ICu + \lambda_\nu(\nu'Iu) - \lambda(\nu'Iu) = 0$$

hence:

$$(\nu'Iu) = \frac{(\lambda\lambda_\nu)^{1/2}\xi\xi_\nu}{\lambda - \lambda_\nu}$$

and therefore:

$$Iu = \sum_{\nu \ne u} \frac{(\lambda\lambda_\nu)^{1/2}\xi\xi_\nu}{\lambda - \lambda_\nu}\nu.$$

Correspondence analysis or multiple correspondence analysis is subject to analogous treatment.

In the case of multiple correspondence analysis (the more general case), each individual is characterized by the vector x_k which "stacks" the indicatrixes of membership in the modalities of p qualitative variables (2 in the case of correspondence analysis). If $\underline{1}$ denotes the vector all of whose components have a value of 1, we have $x_k'\underline{1} = p$ for any k. We then look for vectors u normed by $1/pN\,u'Du = 1$, with $D = \text{diag}\,\sum_U x_k = \sum_U \text{diag}\,x_k$ such that the variance of $\xi_k = 1/p\,x_k'u$ is stationary. This yields a solution to the problem of eigenvalues: $Cu - p\lambda Du = 0$ where $C = \sum_U x_k x_k'$.

The search for a linearized variable for λ and u follows the same procedure as before. We have the relationship between influences:

$$(IC - pI\lambda D - p\lambda ID)u - (C - p\lambda D)\,Iu = 0.$$

As $IC = xx'$, $ID = \text{diag}\,x$, and $u'DIu = 0$, we obtain through premultiplication by u':

$$I\lambda(x) = \frac{1}{N}\left(\left(\frac{x'}{p}u \right)^2 - \lambda\frac{x'}{p}uOu \right)$$

where $uO\nu$ denotes the Hadamard product (*i.e.*, component by component) of u and ν. We know that $u = \underline{1}$ is a eigenvector associated with the eigenvalue 1, also the largest. We check that for $u = \underline{1}$ we have $I\lambda = 0$!

In the same manner, we obtain the components of Iu on the other proper vectors ν:

$$\nu'DIu = \frac{1}{N}\left[\left(\frac{x'}{p}.u \right)\left(\frac{x'}{p}.\nu \right) - \lambda\frac{x'}{p}.uO\nu \right].$$

The analysis may be continued by calculating the variability of a projection onto a factorial design. If A is a subpopulation of size N_A, the coordinates of its representative point on the factorial designs are

$$\alpha_u = \left(\frac{\sum_A x_k}{\sum_A 1} \right)' .u = \bar{X}_A'.u.$$

We linearize α_u by using the relationship $I\alpha_u = (I\bar{X}_A)'u + \bar{X}_A'Iu$ and the rest is simple.

14. CONCLUSION

The linearization of complex statistics has long been considered the most flexible and comprehensive method of obtaining an estimation of the variance. Specifically, this method is applicable to any sample design and to any type of estimator. The popularity of methods based on sample replications is due largely to the fact that certain statistics are considered too complex to be linearized. However, for the large class of substitution estimators, the use of influence functions and of algebraic rules governing their construction makes it possible to obtain fairly simply linearized variables by means of which variance estimation boils down to the estimation of a total estimated using the Horvitz-Thompson estimator.

ACKNOWLEDGEMENTS

I wish to thank Olivier Sautory and Yves Tillé for kindly providing critical comments concerning various drafts of this paper. The latter provided the demonstration at the beginning of section 7. I also wish to thank the two referees of *Survey Methodology*, whose comments did much to make the paper more readable, as well as the editorial board for its understanding.

REFERENCES

BILLINGSLEY, P. (1969). *Convergence of Probability Measures.* New York: Wiley.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A., and KOVAČEVIĆ, M.S. (1997). Variance estimation for measures of income inequality and polarization: The estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys, *Journal of the American Statistical Association*, 89, 1035-1043.

CHAMBERS, R.L., DORFMAN, A.H., and HALL, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, 79, 577-582.

COCHRAN, W. (1977). Sampling Techniques, 3rd edition. New York: Wiley.

DEVILLE, J.-C. (1993). Une formule universelle d'estimation de variance. Internal document, INSEE-UMS.

DEVILLE, J.-C. (1997). Estimation de la variance du coefficient de Gini mesuré par sondage. In *Actes des Journées de Méthodologie Statistiques*, INSEE METHODES, 69-70-71.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

DURBIN, J. (1953). Some results in sampling theory when units are selected with unequal probabilities. *Journal of the Royal Statistical Society B*, 15, 262-269.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 361-374.

HAMPEL, F.R., RONCHETTI, E., ROUSSEEUW, P.J. and STAHEL, W. (1985). *Robust Statistics: The Approach Based on the Influence Function.* New York: Wiley.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

LECOÛTRE, J.P., and TASSI, PH. (1987). Statistique non-paramétrique et robustesse. *Economica.*

ROSEN, B. (1972). Asymptotic theory for successive sampling I and II. *Annals of Mathematical Statistics*, 43, 373-397 and 748-776.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

WOLTER, K.M. (1985). *Introduction to Variance Estimation.* New York: Springer.

WOODRUFF, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 205-212
Statistics Canada

Cosmetic Calibration with Unequal Probability Sampling

K.R.W. BREWER[1]

ABSTRACT

Cosmetic estimators are by definition interpretable both as design-based and as prediction-based estimators. Formulae for them can be obtained directly by equating these two estimators or indirectly by a simple form of calibration. Since they constitute a subset of Generalized Regression Estimators, their design-variances cannot be estimated without knowing the relevant second order inclusion probabilities, but under the prediction model to which they are calibrated those probabilities do not affect their anticipated variances, so it is more appropriate to estimate these and/or their prediction-variances instead. An unanticipated spin-off of cosmetic calibration is a simple and effective method for eliminating negative and unacceptably small positive sample weights. The empirical performance of cosmetically calibrated estimators is put to the test using Australian farm data.

KEY WORDS: Anticipated variance; Design-based estimation; Non-negative weights; Prediction-based estimation; Regression estimation.

1. INTRODUCTION

Cosmetic estimation was introduced by Särndal and Wright (1984). A cosmetic estimator is one that is readily interpretable both as a design-based and as a prediction-based estimator. A procedure for constructing a cosmetic estimator was suggested in Brewer (1995). An improved version of it is presented and discussed in this paper.

Deville and Särndal (1992) described a number of variants on the theme of calibration. The common thread running though them was that for large samples the sample weights had to approximate the Horvitz-Thompson (HT) weights (Horvitz and Thompson 1952); that is to say, the reciprocals, π_j^{-1}, of the first order inclusion probabilities, π_j. Weighted sums of the differences between the calibration weights and the HT weights were minimized to achieve this end. The simplest of Deville and Särndal's calibration estimators was a particular case of the Generalized Regression Estimator or GREG (Cassel, Särndal and Wretman 1976). It requires only a slight modification to become a cosmetic estimator as well. Such estimators will be described here as cosmetically calibrated.

The special case of cosmetic calibration under a stratified simple random sampling design was treated in Brewer (1999). In this paper we consider the generalization to unequal probability sampling. In section 2 the cosmetic and the calibrated approaches to estimation are outlined and shown to be compatible. The design-variance of the cosmetically calibrated estimator is considered in section 3 and its prediction-variance in section 4. It is shown in section 5 that when using cosmetic calibration it is serendipitously easy to overcome the problem of negative and unacceptably small positive weights. Section 6 contains the results of an empirical study based on a somewhat challenging set of Australian farm data. In section 7 the concept is evaluated.

2. THE TWO APPROACHES TO COSMETIC CALIBRATION

Throughout this paper it will be assumed that the population being sampled can be described to a reasonable approximation by the following regression or prediction model:

$$y_j = \mathbf{x}_j'\boldsymbol{\beta} + \varepsilon_j; \quad E_\xi \varepsilon_j = 0,$$

$$E_\xi \varepsilon_j^2 = \sigma^2 a_j^2, \quad E_\xi(\varepsilon_j \varepsilon_k) = 0 \ \forall k \neq j, \tag{1}$$

where $\mathbf{x}_j$ is a p-vector of explanatory variables for unit j and ε_j is a random error with the properties shown, σ^2 is an unknown scalar and the a_j^2 are assumed known. We will also write $\operatorname{diag}(a_j)$ as **A**. Expressions such as "prediction-unbiased" and "prediction-variance" when used in this paper refer to unbiasedness, variance *etc.* in terms of the model (1). It is not uncommonly assumed that the a_j^2 are proportional to some power of a measure of size, say $z_j^{2\gamma}$, where γ lies between 0.5 and 1. When $\gamma = 1$ the coefficient of variation of ε_j is constant. The value $\gamma = 0.5$ corresponds to the situation where the large units behave like random aggregations of small units. Solving the model for the three cases $\gamma = 0.5$, 0.75 and 1 usually gives a realistic range of variance estimates.

The cosmetic approach requires that there be an estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{COS}}$, such that the standard and the predictor forms (Royall 1970) of the GREG estimator are numerically equal, *i.e.*,

$$\hat{T}_{\mathrm{COS}}(\mathbf{y}) = \mathbf{1}_n'\boldsymbol{\Pi}_s^{-1}\mathbf{y}_s + (\mathbf{1}_N'\mathbf{X} - \mathbf{1}_n'\boldsymbol{\Pi}_s^{-1}\mathbf{X}_s)\hat{\boldsymbol{\beta}}_{\mathrm{COS}} \tag{2}$$

$$= \mathbf{1}_n'\mathbf{y}_s + (\mathbf{1}_N'\mathbf{X} - \mathbf{1}_n'\mathbf{X}_s)\hat{\boldsymbol{\beta}}_{\mathrm{COS}}. \tag{3}$$

[1] K.R.W. Brewer, Department of Statistics and Econometrics, Faculty of Economics and Commerce, Australian National University, ACT 0200, Australia.

where $\mathbf{y}$ and $\mathbf{y}_s$ are population and sample vectors of the y values, $\mathbf{X}$ and $\mathbf{X}_s$ are the full-rank $N \times p$ population and $n \times p$ sample matrices respectively of supplementary variables [so that $\mathbf{1}'_N\mathbf{y} = T(\mathbf{y})$ and $\mathbf{1}'_N\mathbf{X} = T(\mathbf{X})$], $\hat{T}_{\mathrm{COS}}(\mathbf{y})$ is the Cosmetic Estimator of $T(\mathbf{y})$ and $\mathbf{\Pi}_s$ is the $n \times n$ diagonal matrix of the sample π_j. It is assumed here that these inclusion probabilities are determined entirely by quantities known to the survey designer from non-sample sources, so that the question of possible informativeness arises only for secondary analysis. $\hat{T}_{\mathrm{COS}}(\mathbf{y})$ also possesses the internally bias-calibrated property defined by Firth and Bennett (1998).

Expressions (2) and (3) are equal when $\mathbf{1}'_n(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)$ $(\mathbf{y}_s = \mathbf{X}_s\hat{\boldsymbol{\beta}}_{\mathrm{COS}}) = 0$. Assuming $\hat{\boldsymbol{\beta}}_{\mathrm{COS}}$ is of the projection form $(\mathbf{Q}'_s\mathbf{X}_s)^{-1}\mathbf{Q}'_s\mathbf{y}_s$, this condition is satisfied when $\mathbf{1}'_n(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)$ spans the row space of $\mathbf{Q}'_s$, for then there must be some row p-vector $\boldsymbol{\alpha}'$ such that $\boldsymbol{\alpha}'\mathbf{Q}'_s = \mathbf{1}'_n(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)$, so $\mathbf{1}'_n(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}_{\mathrm{COS}}) = \boldsymbol{\alpha}'\mathbf{Q}'_s[\mathbf{y}_s - \mathbf{X}_s(\mathbf{Q}'_s\mathbf{X}_s)^{-1}\mathbf{Q}'_s\mathbf{y}_s] = 0$ as required.

Brewer (1995) suggested a way of achieving this result using instrumental variables, but subsequent empirical tests (along the lines of those described in section 6) indicated that this approach was not efficient. It is more efficient, and simpler, to take the Best Linear Unbiased Estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{BLUE}} = (\mathbf{X}'_s\mathbf{A}_s^{-2}\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{A}_s^{-2}\mathbf{y}_s$, where $\mathbf{A}_s$ contains only the sample values in $\mathbf{A}$, and replace the $\mathbf{A}_s^{-2}$ factor by $\mathbf{Z}_s^{-1}(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)$ where $\mathbf{Z}_s$ is $n \times n$ diagonal and $\mathbf{Z}_s\mathbf{1}_n = \mathbf{X}_s\boldsymbol{\alpha}$ is any linear combination of the columns of $\mathbf{X}_s$. For then

$$\hat{\boldsymbol{\beta}}_{\mathrm{COS}} = [\mathbf{X}'_s\mathbf{Z}_s^{-1}(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{X}_s]^{-1}\mathbf{X}'_s\mathbf{Z}_s^{-1}(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{y}_s, \quad (4)$$

which is of the required projection form with $\mathbf{Q}'_s = \mathbf{X}'_s\mathbf{Z}_s^{-1}(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)$. Also, since $\boldsymbol{\alpha}' = \mathbf{I}'_n\mathbf{Z}'_s\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}$, $\boldsymbol{\alpha}'\mathbf{Q}'_s = \mathbf{1}'_n\mathbf{Z}'_s\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{Z}_s^{-1}(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)$. But $\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{Z}_s\mathbf{1}_n = \mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{X}_s\boldsymbol{\alpha} = \mathbf{X}_s\boldsymbol{\alpha} = \mathbf{Z}_s\mathbf{1}_n$, so $\mathbf{1}'_n\mathbf{Z}'_s\mathbf{X}_s(\mathbf{X}'_s\mathbf{X}_s)^{-1}\mathbf{X}'_s = \mathbf{1}'_n\mathbf{Z}'_s$ and $\boldsymbol{\alpha}'\mathbf{Q}'_s = \mathbf{1}'_n(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)$ as required.

The choice of $\mathbf{Z}_s$ is still somewhat arbitrary but, if we aim for the $(\pi_j^{-1} - 1)z_j^{-1}$ to be as closely proportional to the a_j^{-2} as possible, $\hat{\boldsymbol{\beta}}_{\mathrm{COS}}$ can approximate $\hat{\boldsymbol{\beta}}_{\mathrm{BLUE}}$. One case is of particular interest here. If (i) the π_j are chosen to be proportional to the a_j, (aiming to minimize the design-variance of the GREG), (ii) they are all small compared with unity, so that $\pi_j^{-1} - 1 \approx \pi_j^{-1}$, and (iii) the a_j themselves are proportional to the elements $\check{z}_j$ of $\check{\mathbf{z}}$, a linear combination of the columns of $\mathbf{X}$, then the choice $z_j = \check{z}_j$ will achieve the desired close proportionality.

An alternative way to derive the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{COS}}$ is to use the calibration approach described by Deville and Särndal (1992), in which sample weights are made as "close" as possible to the π_j^{-1}, subject to the condition that, for every variable in the columns of $\mathbf{X}$, the sample estimate defined by these weights should be without error. The "closeness" is defined by an arbitrary distance function, but for our present purposes the appropriate function is

$$D = (\mathbf{w}_s - \boldsymbol{\varpi}_s)'[(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}]^{-1}(\mathbf{w}_s - \boldsymbol{\varpi}_s)$$
$$+ 2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{1}_n - \mathbf{X}'_s\mathbf{w}_s),$$

where $\mathbf{w}_s$ is the n-vector of the sample weights w_j, $\boldsymbol{\varpi}_s = \mathbf{\Pi}_s^{-1}\mathbf{1}_n$ is the n-vector of the inverse inclusion probabilities π_j^{-1} and $\boldsymbol{\lambda}'$ is a $1 \times p$ row vector of undetermined multipliers. [This is the same as the first variant of Calibration Estimation used in Deville and Särndal (1992), except that $\mathbf{\Pi}_s^{-1}\mathbf{Z}_s^{-1}$ is replaced by $(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}$.] Differentiating with respect to $\mathbf{w}_s$,

$$\frac{\partial D}{\partial \mathbf{w}_s} = 2[(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}]^{-1}(\mathbf{w}_s - \mathbf{\Pi}_s^{-1}\mathbf{1}_n) - 2\mathbf{X}_s\boldsymbol{\lambda}.$$

Solving $\dfrac{\partial D}{\partial \mathbf{w}_s} = 0$ yields

$$\mathbf{w}_s = \mathbf{\Pi}_s^{-1}\mathbf{1}_n + (\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}\mathbf{X}_s[\mathbf{X}'_s(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}\mathbf{X}_s]^{-1}$$
$$\times (\mathbf{X}'\mathbf{1}_N - \mathbf{X}'_s\mathbf{\Pi}_s^{-1}\mathbf{1}_n), \quad (5)$$

and the corresponding Calibration Estimator, defined as $\mathbf{w}'_s\mathbf{y}_s$, reduces to the formula given for $\hat{T}_{\mathrm{COS}}(\mathbf{y})$ in its GREG form, shown in (2) above. Since (2) and (3) are equivalent, there is also an alternative formula for $\mathbf{w}_s$, namely

$$\mathbf{w}_s = \mathbf{1}_n + (\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}\mathbf{X}_s[\mathbf{X}'_s(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}\mathbf{X}_s]^{-1}$$
$$\times (\mathbf{X}'\mathbf{1}_N - \mathbf{X}'_s\mathbf{1}_n). \quad (6)$$

$\hat{T}_{\mathrm{COS}}(\mathbf{y})$ being the intersection of Särndal and Wright's (1984) Cosmetic Estimators with Deville and Särndal's (1992) Calibration Estimators, we will refer to it from now on as the Cosmetic Calibration Estimator and will write it as $\hat{T}_{\mathrm{COSCAL}}(\mathbf{y})$. Similarly we will write $\hat{\boldsymbol{\beta}}_{\mathrm{COS}}$ as $\hat{\boldsymbol{\beta}}_{\mathrm{COSCAL}}$.

3. DESIGN-VARIANCE AND ANTICIPATED VARIANCE

We consider first the design-variance of $\hat{T}_{\mathrm{HT}}(\mathbf{y})$ and also that of any GREG estimator that is prediction-unbiased under the model (1). Such an estimator can be written $\hat{T}_{\mathrm{GREG}}(\mathbf{y}) = \hat{T}_{\mathrm{HT}}(\mathbf{y}) + \{T(\mathbf{X}) - \hat{T}_{\mathrm{HT}}(\mathbf{X})\}\hat{\boldsymbol{\beta}}_{\mathrm{GREG}}$ where $\hat{\boldsymbol{\beta}}_{\mathrm{GREG}}$ is any prediction-unbiased and prediction-consistent estimator of $\boldsymbol{\beta}$. If the sample size is fixed at n, the design-variance of $\hat{T}_{\mathrm{HT}}(\mathbf{y})$ is

$$V_p\hat{T}_{\mathrm{HT}}(\mathbf{y}) = \sum_{j=2}^{N}\sum_{k=1}^{j-1}(\pi_j\pi_k - \pi_{jk})(y_j\pi_j^{-1} - y_k\pi_k^{-1})^2 \quad (7)$$

where π_{jk} is the joint probability of the inclusion of units j and k in sample. If (1) holds, $\hat{T}_{\mathrm{GREG}}(\mathbf{y}) = T(\mathbf{X})\boldsymbol{\beta} + \hat{T}_{\mathrm{HT}}(\boldsymbol{\varepsilon})$, where $\boldsymbol{\varepsilon}$ is the vector of the ε_j, so writing ε_j in the place of y_j in (7):

$$V_p \hat{T}_{\mathrm{GREG}}(\mathbf{y}) = \sum_{j=2}^{N} \sum_{k=1}^{j-1} (\pi_j \pi_k - \pi_{jk}) \left[\varepsilon_j \pi_j^{-1} - \varepsilon_k \pi_k^{-1} \right]^2. \qquad (8)$$

The design-variances of the HT and the GREG estimators are therefore both functions of the π_{jk}. Important problems that this fact raises have been discussed in some detail by Särndal (1996). Basically they are that the π_{jk} tend to be difficult to evaluate and that their use involves a cumbersome double summation. [To this we may add that the Sen-Yates-Grundy (SYG) variance estimator (Sen 1953, Yates and Grundy 1953), which is usually the most efficient one to use when the sample size is fixed, is easily destabilised by the presence of small values among the π_{jk}, and is biased if any of the π_{jk} are zero. Zero values can occur easily, particularly when sampling systematically. Other relevant references on the π_{jk} include Rao and Bayless (1969), Bayless and Rao (1970) and Brewer and Hanif (1983, 62-68).]

Särndal's proposals were to circumvent the problems, either by relaxing the requirement that the first order inclusion probabilities be exactly proportional to size or the demand that the sample size be fixed. Here we suggest another way to circumvent these problems. It depends on the fact that if the working model (1) holds exactly, then the variations in the π_{jk} from one selection method to another (holding the first-order inclusion probabilities constant) contribute nothing to the prediction-variance of $\hat{T}_{\mathrm{GREG}}(\mathbf{y})$, and hence only trivially to its design-variance. Further, the anticipated variance (AV) of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y}) - T(\mathbf{y})\}$, as defined by Isaki and Fuller (1982), which is its variance under both the design and model (1), is asymptotically independent of the π_{jk}. This may be seen as follows. Since $\hat{T}_{\mathrm{GREG}}(\mathbf{y})$ is both prediction-unbiased and asymptotically design-unbiased (Brewer 1979, Särndal and Wright 1984),

$$AV\{\hat{T}_{\mathrm{GREG}}(\mathbf{y}) - T(\mathbf{y})\}$$

$$\approx E_\xi V_p \hat{T}_{\mathrm{GREG}}(\mathbf{y})$$

$$= \sigma^2 \sum_{j=2}^{N} \sum_{k=1}^{j-1} (\pi_j \pi_k - \pi_{jk}) \left(\pi_j^{-2} a_j^2 + \pi_k^{-2} a_k^2 \right)$$

$$= \frac{1}{2} \sigma^2 \sum_{j=1}^{N} \sum_{\substack{k=1 \\ k \neq j}}^{N} \left(\pi_k \pi_j^{-1} a_j^2 + \pi_j \pi_k^{-1} a_k^2 - \pi_{jk} \pi_j^{-2} a_j^2 - \pi_{jk} \pi_k^{-2} a_k^2 \right)$$

$$= \sigma^2 \sum_{j=1}^{N} \left[(n - \pi_j) \pi_j^{-1} - (n-1) \pi_j^{-1} \right] a_j^2$$

$$= \sigma^2 \sum_{j=1}^{N} \left(\pi_j^{-1} - 1 \right) a_j^2. \qquad (9)$$

This is the same expression as was shown by Godambe (1955) to be the minimum possible anticipated variance (given the values of π_j) for any design-unbiased estimator of $T(\mathbf{y})$. (It also provides the justification for the choice of $\pi_j \propto a_j$ when seeking to minimize the design-variance.) It would therefore seem preferable, if (1) is indeed a useful working model, to estimate the AV of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y}) - T(\mathbf{y})\}$ rather than the design-variance of $\hat{T}_{\mathrm{GREG}}(\mathbf{y})$. It follows immediately from (9) that a large-sample estimator of this AV is $\hat{\sigma}^2 \sum_{j=1} (\pi_j^{-1} - 1) a_j^2$, where $\hat{\sigma}^2$ can be the estimator of σ^2 obtained from a regular regression package based on the use of $\hat{\boldsymbol{\beta}}_{\mathrm{BLUE}}$, but preferably from one in which $\hat{\boldsymbol{\beta}}_{\mathrm{COSCAL}}$ takes the place of $\hat{\boldsymbol{\beta}}_{\mathrm{BLUE}}$ (*cf.* Fuller 1975). Since the only approximation involved in deriving (9) is the omission of terms arising from the design-bias, the proposed estimator may perform reasonably well in smaller samples where the design-bias is known to be small; as, for example, where a regression of the y_j on a single supplementary variable goes almost through the origin.

If, for each assumed value of γ, the sample a_j^2 are normalized to sum to (say) n, the values of $\hat{\sigma}^2$ will be comparable, but the best choice of γ is not necessarily the one that minimizes $\hat{\sigma}^2$. A robust estimator of γ can be obtained by finding the value of $\hat{\gamma}$ for which the correlation between $(y_j - \mathbf{x}_j \hat{\boldsymbol{\beta}})^2 / z_j^{2\hat{\gamma}}$ and the rank of z_j is zero. However, estimates of γ, except where they come from large samples, are typically subject to high variance, and should be treated with caution, especially if they lie outside the range $0.5 \le \gamma \le 1$.

When the analysis is secondary (*i.e.*, not carried out by the person or organization responsible for the design or conduct of the survey) the unavailability of certain relevant information can cause the sample selection to be informative. Special precautions are then usually needed when estimating the model (Pfeffermann, Skinner, Holmes, Goldstein and Rabash 1998). However if the sample values of all the $\mathbf{x}_j$ are known, σ^2 can be estimated using standard regression analysis and the only problem lies in the estimation of the expression $\sum_{j=1}^{N} (\pi_j^{-1} - 1) a_j^2$. The HT estimator $\sum_{j \in s} \pi_j^{-1} (\pi_j^{-1} - 1) a_j^2$ is always available as a last resort, but if a population total such as that of the z_j or of the π_j is known, or better still both are known, that estimator can be improved upon.

4. PREDICTION-VARIANCE

It is appropriate to estimate the anticipated variance for sample design purposes, but for the analysis of any particular sample the prediction-variance is a more logical choice. That prediction-variance is, by definition,

$$E_\xi\left[\hat{T}_{\mathrm{COSCAL}}(\mathbf{y}) - T(\mathbf{y})\right]^2 = E_\xi\left[\sum_{j\in s} w_j y_j - \sum_{j=1}^{N} y_j\right]^2$$

$$= E_\xi\left[\sum_{j\in s} (w_j - 1)y_j - \sum_{j\notin s} y_j\right]^2$$

$$= \sigma^2\left[\sum_{j\in s} (w_j - 1)^2 a_j^2 + \sum_{j\notin s} a_j^2\right]$$

$$= \sigma^2\left[\sum_{j\in s} w_j(w_j - 1)a_j^2 + \left(\sum_{j=1}^{N} a_j^2 - \sum_{j\in s} w_j a_j^2\right)\right] \qquad (10)$$

$$= \sigma^2\left[\sum_{j\in s} w_j(w_j - 1)a_j^2 + \left(\sum_{j=1}^{N} a_j^2 - \sum_{j\in s} \pi_j^{-1} a_j^2\right) - \left(\sum_{j\in s} w_j a_j^2 - \sum_{j\in s} \pi_j^{-1} a_j^2\right)\right]. \qquad (11)$$

Assuming the a_j^2 are known or can be satisfactorily imputed, (10)–like (9)–can be estimated prediction-consistently by replacing σ^2 by $\hat{\sigma}^2$. [However the w_j are not defined for the nonsample units, so it is not possible to use either (10) or (11) to obtain a formula or estimator for the AV of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y}) - T(\mathbf{y})\}$.]

The first expression in round brackets in (11) is the difference between the population sum of the a_j^2 and its HT estimator, and has design expectation zero. Further, since w_j tends asymptotically to π_j^{-1}, the second such expression in (11) is asymptotically zero and negligible for large samples. Hence a simpler but still prediction-cum-design-consistent estimator of the prediction variance of $\{\hat{T}_{\mathrm{GREG}}(\mathbf{y}) - T(\mathbf{y})\}$ is $\hat{\sigma}^2\sum_{j\in s} w_j(w_j - 1)a_j^2$. Since this does not require knowledge of the non-sample a_j^2, it is an attractive choice for secondary analysis.

Both the suggested estimators conveniently take the value zero when every unit in the population is also in sample with $w_j = 1$ for all j. However if the disparity between the population mean and the sample mean is substantial, it may, as in section 3, be necessary to construct special estimators of the unknown $\sum_{j=1}^{N} a_j^2$ and related population sums by calibrating on whatever relevant population data may be available.

5. THE PROBLEM OF NEGATIVE AND OTHER UNACCEPTABLY SMALL SAMPLE WEIGHTS

It was pointed out in Brewer (1999) that strong conditions had to be fulfilled before the Representative Principle underlying design-based inference could be regarded as useful. (This Principle required that for every sample unit included with probability π_j there should be approximately $\pi_j^{-1} - 1$ units with reasonably similar properties in the non-sample portion of the population.) Such strong conditions can nevertheless hold when both the population and the sample are large and the inclusion probabilities are an explicit function of a known measure of size, which is usually a linear function of the columns of **X**.

The manner in which the Cosmetic Calibration weights, w_j, are constructed, however, implies that they are better indexes of the relevant properties than the π_j^{-1} are themselves. So there is a sense in which the inverse weights, w_j^{-1}, can be thought of as analogous to inclusion probabilities. Sample units with large weights (and hence small w_j^{-1}) can be considered as typical in their characteristics in that they "represent" large numbers of population units. Sample units with smaller weights can still be regarded as typical, but they "represent" fewer population units. A sample unit with weight unity is only on the borderline of being typical. It does not represent any other unit. A sample unit with a weight less than one is definitely atypical. It does not even represent itself. A sample unit with a negative weight is perversely atypical and counter-representative. For a small enough domain, it can actually produce negative estimates of total. Its presence in the sample is a "rare event". Yet it must be part of the population, or it could not be in the sample.

The obvious procedure to adopt for a unit with $w_j < 1$ is to delete it both from the sample and the sample frame, to recalculate the w_j so that the remaining sample units are calibrated on the totals of the remaining population units, and then add the deleted unit on as an atypical extra. This, of course, is precisely what many design-oriented survey statisticians have been doing with "outlying observations" for decades. It is also the natural thing to do with sample units that are allocated weights in an unacceptable range.

If, however, we start with a GREG that has not been cosmetically calibrated and attempt to remove the unacceptable weights by setting them at unity and recalculating the remainder, we usually find that many of the newly recalculated weights are themselves unacceptable. If the procedure is taken through further iterations, the number of units whose weights have been set to unity increases steadily, and the larger positive weights that are needed for those that remain leads to a substantial increase in prediction-variance.

This problem can be substantially reduced by using cosmetic calibration. Wherever a sample contains one or more units with such unacceptable weights, each of the corresponding π_j values can (by a convenient fiction) be set equal to one, and the calculation then repeated. The factor $(\mathbf{\Pi}_s^{-1} - \mathbf{I}_n)\mathbf{Z}_s^{-1}$ in (5) and (6) ensures that wherever π_j is set equal to unity, the corresponding w_j is also unity. The comparable factor for the standard GREG, $\mathbf{\Pi}_s^{-1}\mathbf{Z}_s^{-1}$, does not possess this property.

Removing negative and unacceptably small positive weights in this fashion provides no absolute guarantee that the remaining weights do not include some large ones. There is then the danger of introducing a substantial design-bias, but the results of the empirical study presented in the next section suggest that this danger is less than might be feared. Where the inclusion probabilities increase only modestly with size, the cosmetically calibrated GREG can be seen to reduce the incidence of unacceptable weights substantially, and for one sample design entirely eliminate it, without materially increasing the design-variance or introducing any appreciable squared bias term into the design-MSE (mean squared error).

6. AN EMPIRICAL STUDY USING AUSTRALIAN FARM DATA

The actual performance of cosmetic calibration as compared with certain alternative estimation procedures has been studied using data obtained from two farm surveys conducted by the Australian Bureau of Agricultural and Resource Economics using economic and production data collected from a sample of 904 farms in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) and Australian Dairy Industry Survey (ADIS) in the late 1980s (Chambers 1996). The data set includes two variables (incomes from wheat and dairy sales), that follow model (1) reasonably well, and two others (incomes from sheep and beef sales) that do not. These properties make it a useful and exacting data set for testing purposes.

Chambers carried out a comparison of various estimation strategies using three sets of stratified random subsamples from his 904 sample farms. Each set consisted of 500 stratified simple random samples of 100 farms. The size variable used for stratification purposes was Dry Sheep Equivalents (DSEs). For the present study, three additional sets were selected, each again consisting of 500 samples of 100 farms. The inclusion probabilities used in each set were proportional to a fractional power of the DSE. For Set 4 that power was 0.60, for Set 5 it was 0.75, and for Set 6 it was 0.90. In each case there was also a completely enumerated sector. It was smallest for Set 4 and largest for Set 6.

The larger of the two versions of model (1) used by Chambers to construct his estimators had eleven supplementary variables: hectares of wheat, numbers of sheep, beef and dairy cattle, and seven zero-one Industry indicators. This model provided the stronger challenge, and the comparisons presented here relate to that model only.

Chambers calculated sample weights and RMSEs (root mean squared errors) for each of Sets 1-3 using six different estimators. The first of these, "RATIO", was the HT ratio estimator based on each survey variable's natural supplementary variable (such as hectares of wheat for wheat income). He calculated this only as a basis of comparison for other estimators, holding it to be essentially unsatisfactory in that the sample weights differed from one survey variable to another.

The five estimators (other than "RATIO") that were used by Chambers were the standard "GREG" [identical with the first variant of Calibration Estimation used in Deville and Särndal (1992), the variable z in this instance being DSE], "BLUP," (the Best Linear Unbiased Predictor), "RIDGE," (a ridge regression estimator that enabled all weights to fall within the acceptable range) and two estimators, "NWD3" and "NWDAR3," that applied Nadaraya-Watson nonparametric adjustments to the weights for the estimator "RIDGE".

As a supplement to Chambers' study, the cosmetic calibration estimator, "COSCAL," was calculated for Sets 2-6. "COSCAL" and "GREG," having nearly identical formulae, usually had very similar MSEs. Since Chambers' only reason for introducing "RIDGE," "NWD3" and "NWDAR3" was to get rid of unacceptable sample weights, the relevant comparisons in MSE terms are those between these three estimators and "COSCAL."

Except in Set 6, all or nearly all the "COSCAL" weights for any given sample were eventually made greater than or equal to one, but occasionally one or more of the 100 weights could not be found a value in the acceptable range. The most intractable instances occurred where only three farms had been selected for the Dairy Industry and all three of them were of larger than average size. It was therefore logically impossible to calculate any set consisting of all positive weights to specify an estimator calibrated both on the number of dairy farms and on the Dairy Industry's total size measure. (Dairy farms in Australia are typically on the small side, so for a sample of given size there are fewer dairy farms selected when the probability of inclusion increases rapidly with size of farm than when it increases slowly.)

The actual extent of the unacceptable weight problem is indicated in Table 1. The elimination procedure broke down completely only for Set 6. It seems probable that there was less of a problem for Set 3 than for Set 6 because Set 3 had no inclusion probabilities that were close to but not equal to one.

Table 2 shows the initial incidences of unacceptable weights for the estimators "GREG", "BLUP/RIDGE" and "COSCAL". The corresponding final incidence for "BLUP/RIDGE" is uniformly zero, but the estimator is then no longer "BLUP" but "RIDGE". For Set 2 the initial incidence of such small weights is substantially less for "BLUP" than it is for "GREG" or "COSCAL". For Set 3, however, the initial incidence for "COSCAL" is substantially smaller than that for "GREG" or "BLUP". For Set 6 ($\pi_j \propto \mathrm{DSE}_j^{0.90}$) the number of unacceptable weights found and the number of intractable samples discovered were already unacceptably large after only two iterations.

RMSEs were obtained for "COSCAL", both before and after the unacceptable sample weights had been eliminated as far as possible. Table 3 contains a comparison between these RMSEs and those reported in Chambers (1996) for the other estimators.

Most of the RMSEs obtained for the final versions of "COSCAL" are very similar to those obtained from the initial versions, and also from the standard "GREG". The deterioration seen for "COSCAL" in the Dairy Income estimates is due to the small number of Dairy Industry farms selected (particularly in Set 3) and the consequently rapid rise in RMSE that occurred as more and more farms with unacceptable weights were given unit weight and the effective sample size was consequently decreased. The same is true to a lesser extent for wheat farms.

Table 1
Progressive Elimination of Unacceptable
COSCAL Sample Weights

Sample Set	Iteration number	Number of samples with sample weights < 1	Intractable samples detected	Number of sample weights < 1 across samples
Set 2	0	277	0	496
	1	85	0	127
"Compromise"	2	18	0	29
allocation	3	7	0	16
	4	2	2	3
Set 3	0	226	0	701
	1	100	1	303
"Optimal"	2	48	1	134
allocation	3	27	4	75
	4	13	7	48
	5	8	7	39
	6	8	8	39
Set 4	0	188	0	322
	1	55	0	80
Allocation	2	11	0	14
$\propto$ DSE$^{0.60}$	3	0	0	0
Set 5	0	204	0	341
	1	51	0	77
Allocation	2	11	0	16
$\propto$ DSE$^{0.75}$	3	3	1	4
	4	1	1	2
Set 6	0	187	0	592
	1	96	1	229
Allocation	2	46	6	154
$\propto$ DSE$^{0.90}$		Further analysis abandoned		

After the elimination of unacceptable weights, the cosmetically calibrated estimator is design-biased. This is on account of atypical farms that were not selected with certainty being given unit weight. Table 4 shows that for all variables other than Dairy Income the change in the bias between Initial and Final/Intermediate was less than one third of a percentage point. For Dairy Income it is 3.21% for the intractable Set 6, 1.25% for the next most intractable Set 3 and 0.53% or less for the remainder. In every case the squared bias is less than 11% of the MSE, being largest for Wheat Income Set 6 (both Initial and Final).

Table 4 also supplements Table 3 in giving data both on the accuracies of the sample estimates obtained using Sets 4, 5 and 6 and on the percentage Mean Average Deviation Errors (% MADE) for all sets. Sets 4 and 5 seem close to having optimal inclusion probabilities, both in terms of MSE and in the ease with which unacceptable weights can be removed (seemingly just a coincidence, but certainly a happy one). By contrast, Set 6 performs rather poorly. The ratios of the MADEs to the MSEs almost all fall in the range from 0.52 to 0.68. The three exceptionally small ratios, all for Dairy Income, appear to be indicative of occasional large deviations from the mean when the number of dairy farms selected was particularly small.

Table 2
Percentages of Samples Containing Unacceptable
Sample Weights

Sample Set	GREG		BLUP/ RIDDGE		COSCAL Initial		Interm/Final*	
Set 1 (srswor)	77	(4.27)	77	(4.27)	77	(4.27)	n.c.	(n.c.)
Set 2 "Compromise"	53	(1.83)	20	(1.83)	55	(1.79)	0.4	(1.50)
Set 3 "Optimal"	94	(9.73)	93	(5.87)	45	(3.10)	1.6	(4.88)
Set 4 ($\propto$DSE$^{0.60}$)	n.c.		n.c.		38	(1.71)	0.0	–
Set 5 ($\propto$DSE$^{0.75}$)	n.c.		n.c.		41	(1.67)	0.2	(2.00)
Set 6 ($\propto$DSE$^{0.90}$)	n.c.		n.c.		37	(3.17)	n.c.	(n.c.)

n.c. not calculated.
Numbers in parentheses are the average numbers of sample weights less than unity in samples containing at least one such unacceptable sample weight.
*Intermediate for Set 6, Final for all other Sets.

Table 3
RMSEs of the Estimated Population Means of Survey Variables as
Percentages of the Corresponding Population Values
(Chambers' Original Sets Only)

Estimator	Income From: Wheat	Beef	Sheep	Dairy	Total
			Set 1 (srswor)		
RATIO	14.7	28.9	19.1	14.4	16.7
GREG	13.6	26.1	17.0	15.0	17.3
BLUP	13.6	26.1	17.0	15.0	17.3
RIDGE	15.7	23.6	16.0	17.1	15.7
NWD3	15.0	22.1	15.9	17.5	14.6
NWD3AR	14.5	22.4	15.6	17.0	14.7
			Set 2 "Compromise"		
RATIO	10.0	11.6	15.5	19.2	8.3
GREG	9.9	11.9	14.8	20.3	8.4
BLUP	10.8	12.8	14.3	20.5	8.9
RIDGE	13.2	13.0	15.6	23.1	9.8
NWD3	10.5	11.5	14.1	19.8	8.1
NWD3AR	10.5	11.6	14.1	19.7	8.1
COSCAL:					
Initial	9.9	12.1	14.8	20.3	8.4
Final	9.9	12.0	14.8	21.1	8.4
			Set 3 "Optimal"		
RATIO	10.1	10.1	15.9	25.7	7.9
GREG	11.6	11.6	17.4	32.3	8.4
BLUP	11.9	11.1	16.4	32.1	8.0
RIDGE	23.5	9.6	21.3	47.8	11.9
NWD3	12.5	9.1	15.6	30.7	7.3
NWD3AR	12.9	8.9	15.7	31.5	7.3
COSCAL:					
Initial	11.6	11.4	17.6	32.5	8.3
Final	14.6	11.6	18.1	41.4	8.7

Table 4
Performances of Initial and Final (or Intermediate*) COSCAL Estimates

Survey Variable	Sample Set		Inital			Intermediate/Final*	
		% Bias	% RMSE	% MADE	% Bias	% RMSE	% MADE
Wheat Income							
	Set 2 "Compromise"	0.22	9.9	6.4	0.13	9.9	6.4
	Set 3 "Optimal"	0.99	11.6	7.7	0.67	14.6	7.7
	Set 4 ($\propto$DSE$^{0.60}$)	1.83	8.9	6.0	1.79	8.8	6.0
	Set 5 ($\propto$DSE$^{0.75}$)	2.93	9.7	5.7	2.92	9.7	5.7
	Set 6 ($\propto$DSE$^{0.90}$)	3.45	11.0	7.0	3.52	10.8	6.9
Beef Income							
	Set 2 "Compromise"	- 0.01	12.1	8.1	- 0.08	12.0	8.1
	Set 3 "Optimal"	0.50	11.4	7.0	0.25	11.6	7.0
	Set 4 ($\propto$DSE$^{0.60}$)	2.22	13.0	7.4	2.49	11.4	7.3
	Set 5 ($\propto$DSE$^{0.75}$)	2.00	10.4	6.6	1.97	10.4	6.6
	Set 6 ($\propto$DSE$^{0.90}$)	1.55	11.4	6.2	1.71	10.9	6.4
Sheep Income							
	Set 2 "Compromise"	1.05	14.8	9.9	1.09	14.8	9.9
	Set 3 "Optimal"	0.94	17.6	10.8	0.72	18.1	10.9
	Set 4 ($\propto$DSE$^{0.60}$)	- 0.09	13.5	9.0	- 0.04	13.6	9.0
	Set 5 ($\propto$DSE$^{0.75}$)	0.27	14.5	9.8	0.35	14.4	9.8
	Set 6 ($\propto$DSE$^{0.90}$)	1.04	16.9	9.9	1.11	17.2	10.3
Dairy Income							
	Set 2 "Compromise"	- 0.24	20.3	11.4	0.25	21.1	11.4
	Set 3 "Optimal"	1.32	32.5	15.2	2.57	41.4	15.1
	Set 4 ($\propto$DSE$^{0.60}$)	- 0.52	20.2	11.7	- 0.30	20.1	11.7
	Set 5 ($\propto$DSE$^{0.75}$)	- 2.47	20.4	13.4	- 1.94	21.4	13.5
	Set 6 ($\propto$DSE$^{0.90}$)	0.01	29.8	16.3	- 3.20	57.8	16.6
Total Income							
	Set 2 "Compromise"	0.18	8.4	5.4	0.14	8.4	5.4
	Set 3 "Optimal"	0.69	8.3	5.4	0.46	8.7	5.6
	Set 4 ($\propto$DSE$^{0.60}$)	1.75	8.9	4.6	1.92	7.9	4.6
	Set 5 ($\propto$DSE$^{0.75}$)	1.83	7.5	5.0	1.84	7.5	4.9
	Set 6 ($\propto$DSE$^{0.90}$)	1.83	8.5	4.6	1.87	8.3	4.6

* Intermediate for Set 6, Final for all other Sets.

No single estimator out of "RATIO", "GREG", "BLUP" and "COSCAL" has a consistent edge over any of the others on the sole ground of low RMSE. "RIDGE" is generally inferior, as might be expected on account of its prediction-bias, and the two Nadaraya-Watson estimators are generally superior, as might also be expected on account of their nonparametric calibration. However, their superiority is neither compellingly large nor consistent over all variables.

If the choice is restricted to the three simplest estimators capable of producing the same weights for all variables, namely "BLUP", "GREG" and "COSCAL", then all three are comparable in accuracy but "BLUP" and "GREG" are inferior to "COSCAL" in the elimination of unacceptable sample weights. It is true that "COSCAL" was not uniformly successful in eliminating such weights, but the test it faced was exceptionally severe. Eleven explanatory variables were used for samples of size n = 100, the totals of these explanatory variables included several linked pairs (each consisting of a production measure and a count of the number of contributing farms) and for two of the six sample sets the inclusion probabilities increased rapidly with size. Such a stringent combination of requirements should seldom be encountered in normal survey practice. However, especially in circumstances where the explanatory variables include such linked pairs, it would seem prudent to avoid using inclusion probabilities that increase rapidly with size, even at the expense of a moderate departure from the otherwise optimal rule that the π_j should be proportional to the a_j.

7. EVALUATION

It takes little effort to change a standard GREG estimator into a cosmetically calibrated estimator. The matrix $\mathbf{\Pi}_s^{-1}$ in one formula must be replaced by $\mathbf{\Pi}_s^{-1} - \mathbf{I}_n$, and it may also be desirable to replace the existing $\mathbf{Z}_s$ matrix by another choice. The efficiency of the estimator seems to be little changed as a result, but there are several unequivocal advantages.

(i) The estimator is then clearly interpretable as prediction-based as well as design-based.

(ii) Its anticipated variance and its prediction-variance can both be estimated more easily and more efficiently than the design-variance of the standard GREG. [Although these options are also available for any GREG estimator, the most appropriate estimator of β for the purpose of estimating σ^2 is one that is equally relevant to design-based and prediction-based inference. The $\hat{\beta}_{\mathrm{COS}}$ obtained by equating (2) and (3) is such an estimator.]

(iii) Design-based estimation has a tendency to be more reliable for large samples, and prediction-based estimation for small samples and small domains (Brewer 1999). It is not surprising, therefore, that the estimators used for large domains are typically design-based while those for small domains are often purely prediction-based or "synthetic." If the large-domain estimators are cosmetically calibrated, the estimates for their component small domains automatically sum to them without forcing.

(iv) As an unexpected spin-off, the elimination of negative and other unacceptably small weights is streamlined by the use of cosmetic calibration.

ACKNOWLEDGEMENTS

I am indebted to Prof. R.L. Chambers, who provided me with the initial impetus to consider this problem, and to Prof. A.H. Welsh and Drs. P.S. Kott and D.A.Binder for their advice and encouragement along the way. An Associate Editor and two referees provided extended comments on the paper when it was first submitted, which greatly improved both its content and its presentation.

REFERENCES

BAYLESS, D.L., and RAO, J.N.K. (1970). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling (n = 3 or 4). *Journal of the American Statistical Association*, 65, 1645-1667.

BREWER, K.R.W. (1979). A class of robust sample designs for large scale surveys. *Journal of the American Statistical Association*, 74, 911-915.

BREWER, K.R.W. (1995). Combining design-based and model-based inference. Chapter 30 in *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott.). New York: Wiley, 589-606.

BREWER, K.R.W. (1999). Design-based or prediction-based inference? Stratified random vs. stratified balanced sampling. *International Statistical Review*, 67, 35-47.

BREWER, K.R.W., and HANIF, M. (1983). *Sampling With Unequal Probabilities, Lecture Notes in Statistics*. New York: Springer-Verlag, 15.

CASSEL, C-M., SÄRNDAL, C-E., and WRETMAN, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

DEVILLE, J-C., and SÄRNDAL, C-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

FIRTH, D., and BENNETT, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, Series B, 60, 3-21 with discussion on 41-56.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, Series C, 37, 117-132.

GODAMBE, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society*, Series B, 17, 269-278.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

PFEFFERMANN, D., SKINNER, C.J., HOLMES, D.J., GOLDSTEIN, H., and RABASH, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-40 with discussion on pp 41-56.

RAO, J.N.K., and BAYLESS, D.L. (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units. *Journal of the American Statistical Association*, 64, 540-549.

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

SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.

SÄRNDAL, C.-E., and WRIGHT, R.L. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

SEN, A.R. (1953). On the estimate of the variance when sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 18, 52-56.

YATES, F., and GRUNDY P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, Series B, 15, 253-261.

Survey Methodology, December 1999
Vol. 25, No. 2, pp. 213-221
Statistics Canada

The Use of Auxiliary Information in Design-Based Estimation for Domains

VICTOR M. ESTEVAO and CARL-ERIK SÄRNDAL[1]

ABSTRACT

This paper examines some important issues in the use of auxiliary information to produce design-based estimates for domains. We identify three types of design-based estimators and discuss two of these in detail. Both are defined as linear weighted sums of the observed values y_k of the variable of interest. The first is the linear prediction estimator, which is built on a principle of model fitting and good predictions of the unobserved y_k. The second is the uni-weight estimator, which applies the same weight to y_k in the calculation of all estimates for those domains containing unit k. The latter approach has practical advantages for large-scale productions of statistics because it does not require the calculation of different weight systems for the many variables of interest. It is used in Statistics Canada's Generalized Estimation System (GES), which produces point estimates and corresponding design-based variance estimates for any domains. The auxiliary information used to create the weight system determines the precision of the domain estimates. For the uni-weight estimator in particular, a crucial factor for its variance is the level (domain level, population level, or some intermediate level) for which the auxiliary information is known. We define information groups as the subpopulations with known auxiliary totals. These should be as close as possible to the domains of interest in order to produce efficient estimates. We prove that under certain conditions, the variance of the domain total increases monotonically as the information group moves from the domain to the entire population.

KEY WORDS: Design-based domain estimation; Auxiliary information; Level of auxiliary information; Information groups; Prediction estimator; Uni-weight estimator.

1. INTRODUCTION

The estimation for domains of various sizes is an important requirement in the production of statistics for most government surveys. In most statistical agencies, the estimation and the associated measurement of precision rest on design-based principles as far as possible. As Singh, Gambino and Mantel (1994) point out: "Most producers of survey data are accustomed to design estimators and the corresponding design-based inferences. They interpret the data in the context of repeated samples selected using a given probability sampling design, and use estimated design-based cv's (coefficients of variation). Where possible, samples should be designed to produce small area estimates of adequate precision, and sample designs should be fashioned with this in mind. Auxiliary data should be used, where possible, to improve the precision of direct small area estimates."

Let $U = \{1, ..., k, ..., N\}$ denote the survey population. A probability sample s is drawn from U. The inclusion probability and the sampling weight of unit k are denoted by π_k and $a_k = 1/\pi_k$ respectively. Let U_d be a domain of interest. It can be an arbitrary subpopulation $U_d \subseteq U$. The variable of interest is denoted by y, and y_k is its value for unit k. We want to estimate the population total $Y = \sum_U y_k$. In dealing with a domain, U_d, it is convenient to use the domain specific variable y_d, defined as $y_{dk} = y_k$ if $k \in U_d$, and $y_{dk} = 0$ if $k \notin U_d$. Similarly, if x is an auxiliary variable, we have $x_{dk} = x_k$ if $k \in U_d$, and $x_{dk} = 0$ if $k \notin U_d$. Then we can write $Y_d = \sum_{U_d} y_k = \sum_U y_{dk}$ and $\sum_{s_d} x_k = \sum_s x_{dk}$, where $s_d = s \cap U_d$ denotes the part of the sample s that falls in domain U_d.

2. AN EXAMPLE TO INTRODUCE THE ISSUE

The following example illustrates how different levels of auxiliary information can cause large differences in the variance of an estimator of a domain total. Suppose we have a one-dimensional auxiliary variable x which is strictly positive and positively correlated with the variable of interest y. Simple random sampling without replacement (SRSWOR) is used to draw a sample s of size n from U, so $a_k = N/n$ for all $k \in U$. Consider a domain of interest U_d for which we need to estimate $Y_d = \sum_{U_d} y_k$. The following three design-based estimators come to mind.

$$\hat{Y}_{d1} = (X_d / \hat{X}_{d\pi})\hat{Y}_{d\pi} = X_d \hat{R}_d$$

$$\hat{Y}_{d2} = (X / \hat{X}_{\pi})\hat{Y}_{d\pi} = X \hat{R}_{(d)}$$

$$\hat{Y}_{d3} = \hat{Y}_{d\pi}$$

where $\hat{Y}_{d\pi} = (N/n)\sum_s y_{dk}$, $\hat{X}_{d\pi} = (N/n)\sum_s x_{dk}$, $\hat{X}_{\pi} = (N/n)\sum_s x_k = N\bar{x}_s$, $X_d = \sum_U x_{dk}$, $X = \sum_U x_k$, $\hat{R}_d = \sum_s y_{dk} / \sum_s x_{dk}$ and $\hat{R}_{(d)} = \sum_s y_{dk} / \sum_s x_k$. All three estimators are design

[1] Victor M. Estevao and C.-E. Särndal, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada.

consistent. Estimator $\hat{Y}_{d1}$ uses the auxiliary total X_d at the domain level, whereas $\hat{Y}_{d2}$ uses the auxiliary total X at the population level. In practice, this distinction comes into play when the auxiliary information is derived from a source other than the current survey, such as an administrative data source. The total at the domain level is not always available to construct $\hat{Y}_{d1}$, but we often know the total at the population level required for $\hat{Y}_{d2}$. Since $\hat{Y}_{d1}$ uses more detailed information, we intuitively expect that its variance should be smaller than that of $\hat{Y}_{d2}$. Finally, $\hat{Y}_{d3}$ is the Horvitz-Thompson (HT) estimator. Although design-unbiased, it is usually less efficient since auxiliary information is not used.

The variance of $\hat{Y}_{dj}$ for $j = 1, 2, 3$, is approximately

$$V(\hat{Y}_{dj}) = N^2(1/n - 1/N)P_d\, S^2_{yU_d} H_{dj} \tag{2.1}$$

where $P_d = N_d/N$ is the relative size of the domain, $S^2_{yU_d} = \sum_{U_d}(y_k - \bar{y}_{U_d})^2/(N_d - 1)$ is the variance of y in the domain and H_{dj} is the only factor differentiating the three variances. Let $K_d = \mathrm{cv}_{xU_d}/\mathrm{cv}_{yU_d}$ where $\mathrm{cv}_{yU_d} = S_{yU_d}/\bar{y}_{U_d}$ and $\mathrm{cv}_{xU_d} = S_{xU_d}/\bar{x}_{U_d}$ are the coefficients of variation of y and x within the domain, and let $r_d = S_{xyU_d}/S_{xU_d}S_{yU_d}$ denote the corresponding correlation coefficient. Then we have

$$H_{d1} = 1 + K_d^2 - 2r_dK_d$$

$$H_{d2} = 1 - 2P_dM_dr_dK_d + \left\{1 + P_d\left[1 - 2M_d + (\mathrm{cv}_{xU})^2\right]\right\}\Big/\left(\mathrm{cv}_{yU_d}\right)^2$$

$$H_{d3} = 1 + (1 - P_d)\Big/\left(\mathrm{cv}_{yU_d}\right)^2$$

where $M_d = \bar{x}_{U_d}/\bar{x}_U$ and $\mathrm{cv}_{xU} = S_{xU}/\bar{x}_U$. These expressions can be obtained using, for example, Särndal, Swensson and Wretman (1992), Chapters 6 and 9. The terms H_{d1} and H_{d2} follow from the Taylor variance, that is, the variance of the linearized statistic. The expression for H_{d3} follows from the exact HT variance. The approximations $(N_d - 1)/(N - 1) \cong N_d/N = P_d$ and $(N_d - 1)/N_d \cong 1$ were used in all three cases.

Since the variances $V(\hat{Y}_{dj}), j = 1,2,3$, depend on several parameters, it is not so easy to compare them. In Table 1 we compare the three variances for different values of r_d and P_d under the assumption $M_d = \bar{x}_{U_d}/\bar{x}_U = 1$ and $\mathrm{cv}_{yU_d} = \mathrm{cv}_{xU_d} = \mathrm{cv}_{xU} = 1$. Roughly speaking, we assume that y and x have the same variability in the domain and that x has a similar distribution in the domain as in the population. For domains of size $P_d \le 0.5$, Table 1 shows the following.

1. $\hat{Y}_{d1}$ has considerably smaller variance than $\hat{Y}_{d2}$, particularly as the domain size decreases and the correlation increases. It is not surprising that $\hat{Y}_{d2}$ is less efficient than $\hat{Y}_{d1}$. What is surprising is the rapid rate at which this occurs.

2. $\hat{Y}_{d2}$ has only marginally smaller variance than the HT estimator $\hat{Y}_{d3}$. This is particularly striking for smaller domains ($P_d = 0.1$ and 0.3), but even if the domain is as large as half the population ($P_d = 0.5$), $\hat{Y}_{d2}$ is only moderately more efficient than $\hat{Y}_{d3}$ when the correlation is 0.9 or larger.

Table 1
Variance Ratios Comparing $\hat{Y}_{d1}$, $\hat{Y}_{d2}$ and $\hat{Y}_{d3}$ Under SRSWOR; $V_{dj} = V(\hat{Y}_{dj}), j = 1,2,3$; r_d is the Correlation Between x and y in the Domain; $P_d = N_d/N$ is the Relative Domain Size

	Variance Ratio							
	V_{d2}/V_{d1}				V_{d3}/V_{d2}			
	r_d				r_d			
P_d	0.70	0.80	0.90	0.95	0.70	0.80	0.90	0.95
0.1	3.10	4.60	9.10	18.10	1.02	1.03	1.04	1.05
0.3	2.63	3.80	7.30	14.30	1.08	1.12	1.16	1.19
0.5	2.17	3.00	5.50	10.50	1.15	1.25	1.36	1.43
1.0	1.00	1.00	1.00	1.00	1.67	2.50	5.00	10.00

In fact, under conditions other than those of Table 1, it is easy to see from (2.1) that $\hat{Y}_{d2}$ can have a larger variance than $\hat{Y}_{d3}$. This can happen for example, when $M_d = \bar{x}_{U_d}/\bar{x}_U > 1$ and $\mathrm{cv}_{xU_d}/\mathrm{cv}_{xU} < 1$. Therefore, even though $\hat{Y}_{d2}$ uses a highly correlated auxiliary variable, an estimator which does not, namely $\hat{Y}_{d3}$, may be a better choice. The poor performance of $\hat{Y}_{d2}$ may seem disappointing but what counts is not so much the use of a highly correlated auxiliary variable as the level at which we have information about this variable. Estimator $\hat{Y}_{d2}$ uses auxiliary information at the population level, and this is not very efficient for estimation at the domain level.

In a survey where the frame provides a positive measure of size x_k for every $k \in U$, it is possible to calculate an auxiliary total at any level – for the domain, the whole population, or any other subpopulation in between. We can use any of these totals to form a ratio estimator for the domain. The example suggests that the estimator with the auxiliary total at the domain level is better than one with an auxiliary total at a level above the domain. The gain from using the highly correlated auxiliary variable diminishes rapidly as the level of the known auxiliary total moves from the domain to the entire population.

Note that $\hat{Y}_{d1}$ is not the only design consistent estimator that can be constructed with the auxiliary information at the domain level, X_d. Another possibility is $\hat{Y}_{d4} = \hat{Y}_{d\pi} + (X_d - \hat{X}_{d\pi})\hat{R}$, where the slope estimate $\hat{R} = \sum_s y_k/\sum_s x_k$ is based on the entire sample s, not only on the domain part of the sample as in $\hat{Y}_{d1}$. The difference is in the underlying regression model: a common slope for the whole population in the case of $\hat{Y}_{d4}$ and a separate slope for the domain in the case of $\hat{Y}_{d1}$. The variances $V(\hat{Y}_{d1})$ and $V(\hat{Y}_{d4})$ are equal (to the same order of approximation as in (2.1)) if $R = \sum_U y_k/\sum_U x_k = \sum_{U_d} y_k/\sum_{U_d} x_k = R_d$. They will not be very different even when $R_d \ne R$. That is, the choice of model is relatively unimportant. By contrast, the level of auxiliary information leads to considerable differences in the variance.

3. ISSUES RAISED BY THE EXAMPLE

The results in Table 1 raise several issues. In this paper, we examine three of these, in the general case of a multidimensional auxiliary vector x:

1. The ratio estimators $\hat{Y}_{d1}$ and $\hat{Y}_{d2}$ in the example use the same auxiliary variable x. Thus both should benefit from a positive correlation between x and y. But they are different by construction, and they behave very differently, as the example shows. What are the two construction principles, in the general setting with multidimensional auxiliary vectors? Do these two principles yield identical estimators in some situations? These issues are discussed in sections 4, 5 and 6.
2. For a given domain, we define the information group to be the subpopulation for which the x-total is known. In $\hat{Y}_{d1}$, the domain is the information group and for $\hat{Y}_{d2}$ it is the entire population. The example shows that the level of the information group is an important factor for the variance. Are there conditions for which a lower level group will yield a strictly smaller variance than a higher level group? This issue is discussed in section 7.
3. The domain size (number of units in the domain), is another component of auxiliary information. How are the domain sizes incorporated into $\hat{Y}_{d1}$ and $\hat{Y}_{d2}$ in addition to the auxiliary information on x ? This issue is discussed in section 8.

4. CONSTRUCTION BY PREDICTED VALUES OBTAINED BY MODEL FITTING

4.1 The Prediction Argument in Estimation for the Entire Population

Suppose the target of estimation is the entire population total, $Y = \sum_U y_k$. A sample s is drawn, giving unit k the sampling weight $a_k = 1/\pi_k$. The data $\{y_k: k \in s\}$ are observed. For non-sampled units, y_k is unknown but suppose we can find a value μ_k that approximates y_k for all units in the population, even if only rather crudely. Then there is strong incentive to build the estimator by "shifting the origin" of unit k from 0 to μ_k, because the residuals $y_k - \mu_k$ are smaller on average than the y_k values and have smaller design-based variance. Now $Y = \sum_U \mu_k + \sum_U (y_k - \mu_k)$, where the known sum $\sum_U \mu_k$ is the dominant term, and the smaller residual sum $\sum_U (y_k - \mu_k)$ requires estimation. Conceptually, two choices must now be made:

(i) Treating the μ_k as non-random, we must choose an estimator for the residual sum $\sum_U (y_k - \mu_k)$.

(ii) We need to find values μ_k close to the y_k. There are two parts to this choice: (a) the model relating y_k to μ_k, and (b) the technique used to fit this model: (generalized) least squares, GLIM, maximum likelihood or some other alternative.

The usual choice in step (i) is the HT estimator, leading to $\hat{Y} = \sum_U \mu_k + \sum_s a_k (y_k - \mu_k)$. This choice is made by convention and is not optimal. No minimum variance unbiased choice exists. Alternatives are the estimators considered by Raj (1956) and Murthy (1957). Auxiliary information is important in step (ii). Let x be auxiliary vector of dimension $J \geq 1$, and let x_k be its value for unit k. Suppose x_k is on the sampling frame for every $k \in s$. Predicted values $\hat{y}_k$ are obtained from the auxiliary information by fitting a model so that $E_m(y_k | x_k, \beta) = f(x_k | \beta)$, where E_m is the expectation operator under the model $m, f(\bullet | \beta)$ is a specified function, and β is an unknown vector of model parameters. The model is **linear** if $f(x_k \mid \beta) = x_k' \beta$, otherwise **non-linear**.

Using the sample data $\{(y_k, x_k): k \in s\}$, we obtain $\hat{B}$ as an estimate of β. Then we calculate a predicted value $\hat{y}_k = f(x_k | \hat{B})$, for every $k \in U$. This is feasible because x_k is known for all $k \in U$. Using $\hat{y}_k$ and the HT estimator for the residual sum, we have

$$\hat{Y}_{\text{PRED}} = \sum_U \hat{y}_k + \sum_s a_k (y_k - \hat{y}_k). \tag{4.1}$$

This is the Generalized Regression (GREG) estimator. We use the subscript PRED rather than GREG to emphasize that the construction is based on predicted values. It is an asymptotically design unbiased (ADU) estimator, regardless of whether or not the model m is "true". Hence, it is called model assisted as opposed to model based. It is not known how to obtain an optimal (minimum variance unbiased) estimator of Y under the twofold choice (i) and (ii). The model is linear if $f(x_k \mid \beta) = x_k' \beta$, otherwise non-linear.

4.2 Linear Model

The generalized least squares method is usually used to estimate the parameters of the linear model. Find B to minimize $\sum_s a_k (y_k - x_k' B)^2 / c_k$, where the c_k are suitable positive constants. This leads to

$$\hat{B} = T_s^{-1} \sum_s a_k x_k y_k / c_k \tag{4.2}$$

where $T_s = \sum_s a_k x_k x_k' / c_k$. The predicted values are $\hat{y}_k = x_k' \hat{B}$, and the construction principle (4.1) gives the linear GREG estimator,

$$\hat{Y}_{\text{LINPRED}} = X' \hat{B} + \sum_s a_k (y_k - x_k' \hat{B}) = \hat{Y}_\pi + (X - \hat{X}_\pi)' \hat{B} \tag{4.3}$$

where $X = \sum_U x_k$, $\hat{Y}_\pi = \sum_s a_k y_k$ and $\hat{X}_\pi = \sum_s a_k x_k$. The choice of c_k influences the variance but only in a mild way. For some designs, we can find optimal c_k that minimize the Taylor variance of $\hat{Y}_{\text{LINPRED}}$. The specification of x_k should include information from the sample design. For example, consider a stratified SRS design with H strata and sampling fractions $f_h = n_h/N_h$, for $h = 1, \ldots, H$. Then, to obtain the minimum Taylor variance, we take $1/c_k = 1/f_h - 1$ for all k in stratum h and we let $\boldsymbol{x}_k = (\xi_{1k}, \ldots, \xi_{hk}, \ldots, \xi_{Hk}, \boldsymbol{x}'_{0k})'$, where $\xi_{hk} = 1$ if k is in stratum h, otherwise $\xi_{hk} = 0$ and x_{0k} includes all of the other auxiliary variables with known total $\sum_U \boldsymbol{x}_{0k}$. Then, except for the factor $(n_h - 1)/n_h \cong 1$, (4.3) coincides with an asymptotically optimal solution derived in a different manner by Montanari (1987).

4.3 Non-linear Model

In the nonlinear case, the model m can be fitted by GLIM or some other technique. Generalized least squares remains an expedient approach: minimize the weighted sum of squares $\sum_s a_k(y_k - f(\boldsymbol{x}_k|\boldsymbol{\beta}))^2/c_k$ as discussed by Fuller (1996). This produces an estimator $\hat{\boldsymbol{B}}$ of $\boldsymbol{\beta}$. We then calculate the predicted values $\hat{y}_k = f(\boldsymbol{x}_k|\hat{\boldsymbol{B}})$ for all $k \in U$, and the estimator is built, as in the linear case, according to (4.1). In some recent research, the nonlinear case is compared to the linear case. For a categorical variable of interest taking m possible values, Lehtonen and Veijanen (1998) fit a multinomial logistic model $f(\boldsymbol{x}_k|\boldsymbol{\beta}) = P(y_k = i) = \exp(\boldsymbol{x}'_k \boldsymbol{\beta}_i)/\sum_{r=1}^{m} \exp(\boldsymbol{x}'_k \boldsymbol{\beta}_r)$ for $i = 1, \ldots, m$. They use weighted log-likelihood to estimate $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_m)'$ and from the resulting predicted values $\hat{y}_k$ they build the estimator as in (4.1). Their empirical investigations indicate that this estimator realizes modest efficiency gains compared to the linear GREG estimator. Unlike the linear fit, the $\hat{y}_k$ for the multinomial logistic model are guaranteed to fall in the unit interval. This model is more realistic and provides better fit for many survey data. However, it requires more detailed auxiliary information since the auxiliary values $\boldsymbol{x}_k$ must be known individually for all k. Unlike the linear model, it is not sufficient to simply know the population total of $\boldsymbol{x}_k$ at some level of aggregation. Firth and Bennett (1997) examine the fitting of GLIM models, producing predicted values $\hat{y}_k = G^{-1}(\hat{B}_1 x_{1k} + \ldots + \hat{B}_J x_{Jk})$ where $G(\bullet)$ is the link function. In an empirical study involving tax auditing data, a binary y-variable, and maximum likelihood fit of the simple logistic function, they find that the improvement over the linear GREG estimator is at most, a few percent. These differences are insignificant compared to the very large effects in Table 1 caused by the level of the auxiliary information.

From a variance perspective, it is important that the model m fits well, because the variance depends on the size of the squared residuals $(y_k - \hat{y}_k)^2$. Lehtonen and Veijanen (1998), and Firth and Bennett (1997) show that when we estimate the entire population total Y, there is only a modest decrease in variance in fitting a non-linear model over a linear model. However, in the estimation of a domain total, described below, this decrease in variance is more pronounced. This seems to be especially true as the domain gets smaller, as suggested by the study of Lehtonen and Veijanen (1998).

4.4 The Prediction Argument in Estimating for a Domain

Most surveys require estimates for a large number of domains. There are two simple techniques for constructing a design-based domain estimator. Both take the GREG estimator $\hat{Y}_{\text{PRED}}$ given by (4.1) as the starting point, but the resulting domain estimators are not in general identical. In this section, we present the **predictive argument**, leading up to the **predictive estimator** of a domain total. In section 5.3, we present the **unique weighting argument** leading to the **uni-weight estimator** of a domain total.

The predictive domain estimator is constructed as follows: predicted values $\hat{y}_k$ have been determined for the entire finite population U, under some appropriate model. When the target is a specified domain, it is natural to use the $\hat{y}_k$ and the residuals $(y_k - \hat{y}_k)$ from within that domain only. Replacing U by U_d and s by s_d in (4.1) we get

$$\hat{Y}_{\text{dPRED}} = \sum_{U_d} \hat{y}_k + \sum_{s_d} a_k (y_k - \hat{y}_k). \tag{4.4}$$

By construction, estimator (4.4) meets the objective of additivity over a set of domains that partition the population. That is, the sum of the estimators (4.4) over a set of these domains is equal to $\hat{Y}_{\text{PRED}}$, the entire population estimator given by (4.1). In the special case of linear prediction, $\hat{y}_k = \boldsymbol{x}'_k \hat{\boldsymbol{B}}$, where $\hat{\boldsymbol{B}}$ is given by (4.2). Then (4.4) becomes (see Särndal *et al.*, 1992, Ch. 9)

$$\hat{Y}_{d\text{LINPRED}} = \boldsymbol{X}'_d \hat{\boldsymbol{B}} + \sum_{s_d} a_k (y_k - \boldsymbol{x}'_k \hat{\boldsymbol{B}}). \tag{4.5}$$

The construction of (4.5) requires that the total $\boldsymbol{X}_d = \sum_{U_d} \boldsymbol{x}_k$ be known, either from an exterior source, or from the microdata as when both $\boldsymbol{x}_k$ and domain membership are specified in the sampling frame. If the model produces accurate predictions, $\hat{Y}_{d\text{LINPRED}}$ can be very precise.

Estimators for domains are frequently classified as either **direct** or **indirect**. In the terminology of Schaible (1992) and Federal Committee on Statistical Methodology (1993), an estimator for a domain is called direct only if it uses values of the variable of interest over the domain and for the time period in question. Otherwise, it is indirect. It follows that $\hat{Y}_{d\text{LINPRED}}$ is indirect when $\hat{\boldsymbol{B}}$ is based in part on y-data from outside the domain. For some structures of the auxiliary vector $\boldsymbol{x}_k$, the expression (4.5) for $\hat{Y}_{d\text{LINPRED}}$ is of the direct variety, requiring only $\{y_k : k \in s_d\}$. An example is shown in section 6.

5. CONSTRUCTION VIA A SUPPLY OF WEIGHTS

5.1 The Linear GREG Estimator as a Weighting Procedure

The linear form of $\hat{Y}_{\text{LINPRED}}$ given by (4.3), invites an alternative view. We can regard $\hat{Y}_{\text{LINPRED}}$ as a linear weighted sum of the observed y_k values. This gives

$$\hat{Y}_{\mathrm{LINPRED}} = \sum\nolimits_s w_k y_k \tag{5.1}$$

where $w_k = a_k g_k$, with

$$g_k = 1 + \lambda' \boldsymbol{x}_k / c_k \tag{5.2}$$

and $\lambda' = (\boldsymbol{X} - \hat{\boldsymbol{X}}_\pi)' \boldsymbol{T}_s^{-1}$ where $\boldsymbol{T}_s = \sum_s a_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k$. Note that the g_k depend on both k and s (hence they are random) and their calculation requires the auxiliary total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$. The domain prediction estimator $\hat{Y}_{d\mathrm{LINPRED}}$ given by (4.5) can also be expressed by linear weighting as

$$\hat{Y}_{d\mathrm{LINPRED}} = \sum\nolimits_s w_{dk} y_k \tag{5.3}$$

where the weight is now $w_{dk} = a_k g_{dk}$ with $g_{dk} = \delta_{dk} + \lambda_d' \boldsymbol{x}_k / c_k$ where $\delta_{dk} = 1$ if $k \in s_d$, $\delta_{dk} = 0$ if $k \in s - s_d$ and $\lambda_d' = (\boldsymbol{X}_d - \hat{\boldsymbol{X}}_{d\pi})' \boldsymbol{T}_s^{-1}$. We note two features of the weights w_{dk} in (5.3): (i) each domain requires a separate weight system, because g_{dk} depends not only on k and s but also on the domain U_d; (ii) all units in the sample s (not just those in the domain of interest) may receive non-zero weights g_{dk}. Hence, the estimation may be indirect. Exceptions to this occur when $\boldsymbol{x}_k$ has a structure such that $g_{dk} = 0$ for $k \in s - s_d$. In this case, (5.3) becomes a direct estimator. An example is given in theorem 6.1 of section 6.

5.2 The Case for a Unique Set of Weights

The use of separate weight systems for different domains, as in $\hat{Y}_{d\mathrm{LINPRED}}$ given by (5.3), is usually efficient but may be considered a drawback in large-scale production of statistics. Most government surveys require estimates for many y-variables and for each domain of interest we need to create a domain specific variable y_d as defined in section 1. Timeliness in the dissemination of survey results is important. Estimates must be produced routinely and rapidly for all the y_d-variables, including the corresponding estimates of variance. It may not be practical to produce and manage separate weight systems for every y_d-variable. These factors speak in favour of a multi-purpose weight system that can be applied, with good results, to all y_d-variables.

5.3 The Uni-weight System Estimator for Domains

The weights in (5.1) are $w_k = a_k g_k$, where the g_k, given by (5.2), depend on the $\boldsymbol{x}_k$ for the sampled units and on the known total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ but not on the y_k values. They can be computed once and then applied to any domain specific y_d-variable. The information carried by this weighting system is based on the known total $\boldsymbol{X}$ which we assume includes all the information that is available or that we wish to use. Let us apply the weight system $w_k = a_k g_k$ to the data for domain U_d. We obtain the uni-weight estimator of the domain total Y_d defined by

$$\hat{Y}_{d\mathrm{WEIT}} = \sum\nolimits_{s_d} w_k y_k. \tag{5.4}$$

This is a direct estimator, because it uses y-data only from within the domain. By construction, it is additive over mutually exclusive and exhaustive domains since $\sum_{d=1}^{D} \hat{Y}_{d\mathrm{WEIT}} = \sum_s w_k y_k$. Alternatively, we can write $\hat{Y}_{d\mathrm{WEIT}}$ as $\sum_{s_d} w_k y_k = \sum_s w_k y_{dk}$. This permits us to determine the basic statistical properties (asymptotic unbiasedness and variance) of $\hat{Y}_{d\mathrm{WEIT}}$ from the entire sample s. The subscript dWEIT in (5.4) emphasizes the construction in terms of weighted y-values. If the domain is the entire population U, then the linear prediction estimator (4.5) and the uni-weight estimator (5.4) are identical. Both are equal to the linear GREG estimator (4.3). In general however, they differ for a domain that is a proper subset of U.

The idea of a uni-weight system is the basis for the methodology in GES as given by Hidiroglou (1991) and Estevao, Hidiroglou and Särndal (1995). A single weight system creates economies of scale in many surveys. The uni-weight system is not the most efficient for each of the y_d-variables, but this simple approach can often be used to provide good estimates on a timely basis. Sometimes, there is little to be gained by searching for an "optimal" weighting scheme for each of the y_d-variables. Furthermore, when the $\boldsymbol{x}_k$ values are not available for all units on the frame, there is no choice but to compute the weight system using whatever totals are known from administrative sources. Table 2 provides a summary of the features of $\hat{Y}_{\mathrm{PRED}}$, $\hat{Y}_{d\mathrm{LINPRED}}$ and $\hat{Y}_{d\mathrm{WEIT}}$.

Table 2
Comparison of Non-Linear $\hat{Y}_{d\mathrm{PRED}}$,
Linear $\hat{Y}_{d\mathrm{PRED}}$ ($\hat{Y}_{d\mathrm{LINPRED}}$) and $\hat{Y}_{d\mathrm{WEIT}}$

	Estimator		
	Non-linear $\hat{Y}_{d\mathrm{PRED}}$	$\hat{Y}_{d\mathrm{LINPRED}}$	$\hat{Y}_{d\mathrm{WEIT}}$
Auxiliary information requirement	$\boldsymbol{x}_k$ for all $k \in U$	$\sum_{U_d} \boldsymbol{x}_k$	$\sum_U \boldsymbol{x}_k$
Linear weights (for y_k)	No	Yes	Yes
Uni-weights (for all y_d- variables)	No	No	Yes
Additivity over domains that partition U	Yes	Yes	Yes

5.4 Calibration as a Procedure for Creating Weights

Calibration is a computational procedure designed to produce a system of weights based on the known total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$. The calibration procedure starts with the basic weights $a_k = 1/\pi_k$ and modifies them through the use of auxiliary information $\boldsymbol{X}$. We define and minimize a measure of distance between the original weights a_k and the new weights w_k, subject to the calibration constraint $\sum_s w_k \boldsymbol{x}_k = \boldsymbol{X}$ which states that the new weight system must produce an exact estimate of the known vector total $\boldsymbol{X}$. When the distance function is defined as the generalized least squares measure $\sum_s c_k (w_k - a_k)^2 / a_k$, we get the weights $w_k = a_k g_k$ of $\hat{Y}_{\mathrm{LINPRED}} = \boldsymbol{X}'\hat{\boldsymbol{B}} + \sum_s a_k(y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}})$. GES uses a procedure that permits individual bounding of weights, see Estevao (1994). The function is minimized subject to the calibration constraint and the bounding constraint $u_k \le w_k \le l_k$ for $k \in s$, where the u_k and the l_k are the user specified bounds. This avoids negative weights and large positive weights whenever a solution exists. Several alternative distance measures have been considered and recently evaluated by Stukel, Hidiroglou and Särndal (1996), and Singh and Mohl (1996). Each distance measure produces a slightly different weight system, but these do not generally lead to significantly different point estimates.

5.5 Information Groups

The efficiency of the uni-weight system depends on the availability of auxiliary information for suitable population groups. In general, totals known for smaller groups produce more efficient $\hat{Y}_{d\mathrm{WEIT}}$ estimates than totals known for larger groups. This is illustrated by theorem 7.1. We are led to examine the structure of the auxiliary vector total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ used to compute the uni-weight system $\{w_k = a_k g_k : k \in s\}$ for $\hat{Y}_{d\mathrm{WEIT}}$. The vector total $\boldsymbol{X}$ includes known totals of one or more x-variables either for the whole population U or for a set of population subgroups. The x-variables form a vector that we denote by $\boldsymbol{x}_0$ and call the core vector. We define an information group as a subpopulation with a known core vector total. The groups establish the level of the auxiliary information. The efficiency of the uni-weight system is a function of the core vector variables and the level of the information groups.

Classical post-strata are information groups with $\boldsymbol{x}_{0k} = 1$ for all k. As another example, consider a business survey where $\boldsymbol{x}_{0k} = (x_{1k}, x_{2k})$ is the value for enterprise k of a two-dimensional core vector, where x_{1k} = Number of Employees and x_{2k} = Gross Business Income. If the estimation is based on $\boldsymbol{X} = (\sum_U x_{1k}, \sum_U x_{2k})$, then the information for the core vector is at the entire population level. If $\boldsymbol{X} = (\sum_{U_1} x_{1k}, \sum_{U_1} x_{2k}, \sum_{U_2} x_{1k}, \sum_{U_2} x_{2k})$, then the information about the core vector is at the more disaggregated level defined by U_1 and U_2, for example, a geographical subdivision of the population.

Information groups and domains of interest are different concepts. An information group may be a domain, but in general, domains will cut across information groups. In the above example, the domains of interest may be industry classes. In business surveys, some units change classification. As a result, information groups based on the frame information may not be the same as the domains of interest.

If the core vector has dimension $J \ge 1$, and there are $P \ge 1$ groups, then the auxiliary vector $\boldsymbol{x}_k$ has dimension $J \times P$ and is given by $\boldsymbol{x}_k = (\gamma_{1k}\boldsymbol{x}_{0k}', \ldots, \gamma_{pk}\boldsymbol{x}_{0k}', \ldots, \gamma_{Pk}\boldsymbol{x}_{0k}')'$ where γ_{pk} is the group identifier such that $\gamma_{pk} = 1$ if unit k is a member of group p and $\gamma_{pk} = 0$ otherwise. The vector total that must be known is $\boldsymbol{X}' = \sum_U \boldsymbol{x}_k' = (\boldsymbol{X}_{01}', \ldots, \boldsymbol{X}_{0p}', \ldots, \boldsymbol{X}_{0P}')$, where $\boldsymbol{X}_{0p} = \sum_{U_p} \boldsymbol{x}_{0k}$ is the known core vector total for information group p. In the special case where the information groups form a partition of U, Estevao, Hidiroglou and Särndal (1995) point out that the factors g_k given by (5.2), can be computed group by group. The factors for units in group p will depend on the known total for that group, $\boldsymbol{X}_{0p}$, but not on the other $P - 1$ known totals. Letting $\boldsymbol{T}_{s_p} = \sum_{s_p} a_k \boldsymbol{x}_{0k}\boldsymbol{x}_{0k}'/c_k$ and $\hat{\boldsymbol{X}}_{0p\pi} = \sum_{s_p} a_k \boldsymbol{x}_{0k}$, we have from (5.2), for units k in group p,

$$g_k = 1 + (\boldsymbol{X}_{0p} - \hat{\boldsymbol{X}}_{0p\pi})' \boldsymbol{T}_{s_p}^{-1} \boldsymbol{x}_{0k}/c_k. \qquad (5.5)$$

6. EQUIVALENCE OF THE PREDICTION ESTIMATOR AND THE UNI-WEIGHT ESTIMATOR FOR PARTICULAR CASES

In sections 4 and 5, we examined two arguments for constructing a linear weighted estimator of a domain total. They lead to two possibly different estimators of a domain total, the linear prediction estimator $\hat{Y}_{d\mathrm{LINPRED}}$ given by (4.5) and the uni-weight estimator $\hat{Y}_{d\mathrm{WEIT}}$ given by (5.4). The motivating factor in the prediction approach is to obtain close predictions $\hat{y}_k = f(\boldsymbol{x}_k | \hat{\boldsymbol{B}})$ of the y_k. In the uni-weight approach, we apply one set of weights to all the y_d-variables in the survey. The motivation here is to construct a unique weight system that uses auxiliary information to the fullest extent possible. Model fitting and getting the closest possible predictions are not the primary concerns.

It is important to emphasize the distinction between the auxiliary vector and the amount of auxiliary information used in the estimation. Both $\hat{Y}_{d\mathrm{LINPRED}}$ and $\hat{Y}_{d\mathrm{WEIT}}$ use the same auxiliary vector $\boldsymbol{x}_k$. However, the amount of auxiliary information is not necessarily the same: $\hat{Y}_{d\mathrm{LINPRED}}$ requires $\sum_{U_d} \boldsymbol{x}_k$, a total at the domain level, whereas $\hat{Y}_{d\mathrm{WEIT}}$ requires $\sum_U \boldsymbol{x}_k$, a total at the population level. The estimators $\hat{Y}_{d\mathrm{LINPRED}}$ and $\hat{Y}_{d\mathrm{WEIT}}$ are not in general identical, but they are equal for certain structures of the auxiliary vector $\boldsymbol{x}_k$ as we now show.

We consider D domains, $U_1, \ldots, U_d, \ldots U_D$, forming a partition of U. Let $\delta_k = (\delta_{1k}, \ldots, \delta_{dk}, \ldots, \delta_{Dk})'$ be the domain identifier vector for unit k and $\boldsymbol{x}_k = (\delta_{1k}\boldsymbol{x}_{0k}', \ldots, \delta_{dk}\boldsymbol{x}_{0k}', \ldots, \delta_{Dk}\boldsymbol{x}_{0k}')'$ where $\boldsymbol{x}_{0k}$ is the known core vector for unit k. Then the requirement for $\hat{Y}_{d\mathrm{WEIT}}$, "$\sum_U \boldsymbol{x}_k$ must be known", is equivalent to "$\sum_{U_d} \boldsymbol{x}_{0k}$ must be known for each domain". Next, the requirement for $\hat{Y}_{d\mathrm{LINPRED}}$, "$\sum_{U_d} \boldsymbol{x}_k$ must be known" is equivalent to "$\sum_{U_d} \boldsymbol{x}_{0k}$ must be known", since the domains are non-overlapping. Thus, when we estimate for all D domains, the use of $\hat{Y}_{d\mathrm{LINPRED}}$ requires that "$\sum_{U_d} \boldsymbol{x}_{0k}$ must be known for each domain." Both approaches require the D core vector totals $\sum_{U_d} \boldsymbol{x}_{0k}$, $d = 1, \ldots, D$. Each domain is an information group for the core vector $\boldsymbol{x}_0$. Are $\hat{Y}_{d\mathrm{LINPRED}}$ and $\hat{Y}_{d\mathrm{WEIT}}$ identical in this situation? Although they use the same information, this can not be taken for granted because they are constructed differently. However, the following statement shows that they are in fact identical.

Theorem 6.1 Let $U_1, ..., U_d, ..., U_D$ be domains that form a partition of U. Let the auxiliary vector be $\boldsymbol{x}_k = (\delta_{1k}\boldsymbol{x}'_{0k}, ..., \delta_{dk}\boldsymbol{x}'_{0k}, ..., \delta_{Dk}\boldsymbol{x}'_{0k})'$ such that $\boldsymbol{X}_{0d} = \sum_{U_d}\boldsymbol{x}_{0k}$ is a known core vector total for $d = 1, ..., D$. Then the prediction estimator $\hat{Y}_{d\text{LINPRED}}$ given by (5.3) and the uni-weight estimator $\hat{Y}_{d\text{WEIT}}$ given by (5.4) are identical, and $\hat{Y}_{d\text{LINPRED}} = \hat{Y}_{d\text{WEIT}} = \sum_{s_d} a_k g^{*}_{dk} y_k$ where $g^{*}_{dk} = 1 + (\boldsymbol{X}_{0d} - \hat{\boldsymbol{X}}_{0d\pi})'$ $(\boldsymbol{X}_{0d} - \hat{\boldsymbol{X}}_{0d\pi})' \boldsymbol{T}_{s_d}^{-1}\boldsymbol{x}_{0k}/c_k$ with $\hat{\boldsymbol{X}}_{0d\pi} = \sum_{s_d} a_k \boldsymbol{x}_{0k}$ and $\boldsymbol{T}_{s_d} = \sum_{s_d} a_k \boldsymbol{x}_{0k}\boldsymbol{x}'_{0k}/c_k$.

The proof follows by showing $g_{dk} = g_k = g^{*}_{dk}$ for $k \in s_d$ and $g_{dk} = 0$ for all $k \in s - s_d$. The details are omitted. Theorem 6.1 suggests that when possible, we should determine the uni-weight system by using an important set of domains as information groups. The theorem does not in general hold when the domains overlap. For example, consider two domains U_1 and U_2 with a non-empty intersection, $U_{12} = U_1 \cap U_2$. Let $\boldsymbol{x}_k = (\delta_{1k}x_k, \delta_{2k}x_k)'$. Then to estimate the first domain total, $Y_1 = \sum_{U_1} y_k$, $\hat{Y}_{1\text{WEIT}}$ requires the auxiliary information $\sum_U \boldsymbol{x}_k = (\sum_{U_1} x_k, \sum_{U_2} x_k)'$, while $\hat{Y}_{1\text{LINPRED}}$ requires the more detailed information $\sum_{U_1}\boldsymbol{x}_k = (\sum_{U_1} x_k, \sum_{U_{12}} x_k)'$. The two estimators are not identical.

7. A MONOTONIC PROPERTY OF THE TAYLOR VARIANCE

The level of the information groups will greatly influence the variance of the uni-weight estimator. The example in section 2 illustrated this. Loosely speaking, the closer a domain is to an information group, the smaller the variance. For specific cases, it can be shown that the variance is a monotonic function of the information group. Theorem 7.1 illustrates this.

Consider a domain of interest, U_d, of the sampled population U, such that U_d is wholly contained in a given information group, U_1, with known auxiliary total $\boldsymbol{X}_{01} = \sum_{U_1}\boldsymbol{x}_{0k}$. Consider also an alternative, larger information group, U_2, with known auxiliary total $\boldsymbol{X}_{02} = \sum_{U_2}\boldsymbol{x}_{0k}$, and such that $U_d \subseteq U_1 \subseteq U_2 \subseteq U$. Thus U_1 provides information at a more detailed level than U_2, and U_1 is "closer" than U_2 to the domain of interest U_d. The uni-weight estimator of Y_d based on the group $U_j, j = 1$ or 2, is given by

$$\hat{Y}_{d\text{WEIT}j} = \sum_s a_k g_{kj} y_{dk} \qquad (7.1)$$

where g_{kj} is given by the right hand side of (5.5) if we replace the index p by j. Theorem 7.1 deals with two designs, Poisson sampling (with inclusion probability $\pi_k \propto z_k$, where z_k is a measure of size) and stratified simple random sampling (STSRS). Note that the choice of the c_k is important to obtain the result. In the theorem, $V(\hat{Y}_{d\text{WEIT}j})$ denotes the Taylor variance of $\hat{Y}_{d\text{WEIT}j}$.

Theorem 7.1. Let $U_1 \subseteq U_2$ be any information groups such that $U_d \subseteq U_1 \subseteq U_2 \subseteq U$, where U_d is the domain of interest and $\boldsymbol{X}_{0j} = \sum_{U_j}\boldsymbol{x}_{0k}$ is a known core auxiliary total for $U_j, j = 1, 2$. Then the following holds: (a) under Poisson sampling, $V(\hat{Y}_{d\text{WEIT}1}) \le V(\hat{Y}_{d\text{WEIT}2})$ provided $c_k = 1/(a_k - 1)$; (b) under STSRS $V(\hat{Y}_{d\text{WEIT}1}) \le V(\hat{Y}_{d\text{WEIT}2})$ if $\boldsymbol{x}_{0k}$ is defined to include, in addition to other auxiliary variables, the stratum identifier $(\delta_{1k}, ..., \delta_{hk}, ..., \delta_{Hk})$ where $\delta_{hk} = 1$ if unit k belongs to stratum h and $\delta_{hk} = 0$ otherwise, and $c_k = 1$ for all k.

The proof of (a) is given in Appendix A. The proof of (b) is similar. In practice, we usually settle (or have to settle) for a single set of information groups and calculate the uni-weight system as a consequence of this choice. Theorem 7.1 requires rather special conditions but it suggests that in general, we can obtain efficient estimates for important domains by using them as information groups. However, other domains if interest may cut across these information groups and for these domains, the conditions for precise estimates may not be as favourable.

8. INCORPORATING INFORMATION ABOUT DOMAIN SIZES

In this section, we return to the example of section 2 with more auxiliary information. In addition to the unidimensional, always positive core auxiliary variable x with a total known either at the population level or at the domain level, we assume now that there is also information about the sizes, $N_d, d = 1, ..., D$, for D domains of interest forming a partition of U.

By formulating the auxiliary vector x_k as in Case A or Case B below, we can incorporate the known domain sizes into the estimator through either of the two construction arguments in sections 4.4 and 5.3. We use the following notation: $S_{xyd} = \sum_{s_d} a_k (x_k - \tilde{x}_{s_d})(y_k - \tilde{y}_{s_d})$, $S_{xxd} = \sum_{s_d} a_k (x_k - \tilde{x}_{s_d})^2$, $S_{yyd} = \sum_{s_d} a_k (y_k - \tilde{y}_{s_d})^2$, where $\tilde{x}_{s_d} = \sum_{s_d} a_k x_k / \hat{N}_d$, $\tilde{y}_{s_d} = \sum_{s_d} a_k y_k / \hat{N}_d$, and $\hat{N}_d = \sum_{s_d} a_k$. We also use the domain identifier $\delta_k = (\delta_{1k}, ..., \delta_{dk}, ..., \delta_{Dk})'$.

Case A Specify the auxiliary vector as $\boldsymbol{x}_k = (\delta'_k, x_k)'$ and let $c_k = 1$ for all k. Then, the predictive estimator (Case A1) and the uni-weight estimator (Case A2) are not identical.

Case A1 Jointly, the set of D predictive domain estimators (4.5) requires the information $(N_d, X_d), d = 1, ..., D$, where $X_d = N_d \bar{x}_{U_d} = \sum_{U_d} x_k$ is a known total at the domain level. The prediction estimator for domain U_d becomes

$$\hat{Y}_{d\text{LINPRED}} = N_d \tilde{y}_{s_d} + \hat{B}_{\text{COMB}}(X_d - N_d \tilde{x}_{s_d}) \qquad (8.1)$$

where $\hat{B}_{\text{COMB}} = \sum_{d=1}^{D} S_{xyd} / \sum_{d=1}^{D} S_{xxd}$. The underlying regression has a common regression slope for all D domains, whereas the intercepts are allowed to vary between domains. This estimator is indirect since the sampled units in all domains (and not only those in U_d) receive non-zero weights. Each domain has its own weight system.

Case A2 Jointly, the set of uni-weight estimators (5.4) requires the information $(N_1, ..., N_d, ..., N_D, X)$, where $X = \sum_U x_k$ is a known total at the population level. The uni-weight estimator for domain U_d becomes

$$\hat{Y}_{d\text{WEIT}} = N_d \tilde{y}_{s_d} + \hat{B}_{(d)} (X - \textstyle\sum_{d=1}^{D} N_d \tilde{x}_{s_d}) \tag{8.2}$$

where $\hat{B}_{(d)} = S_{xyd} / \sum_{d=1}^{D} S_{xxd}$. This direct estimator is the one that we would have to use if the x-total is not available at any level lower than the entire population. The predictive estimator (8.1) will ordinarily have much lower variance than the uni-weight estimator (8.2). Although they are both based on the same auxiliary vector, the information content is higher for (8.1).

Case B Specify the auxiliary vector as $\boldsymbol{x}_k = (\delta_{1k}\boldsymbol{x}_{0k}', ..., \delta_{dk}\boldsymbol{x}_{0k}', ..., \delta_{Dk}\boldsymbol{x}_{0k}')'$ and let $c_k = 1$ for all k. Here the core vector is $\boldsymbol{x}_{0k} = (1, x_k)'$, and each domain is an information group. Because $\boldsymbol{x}_k$ has this structure, theorem 6.1 tells us that the prediction estimator and the uni-weight estimator are identical. By either approach, the required information for the D estimates is (N_d, X_d), $d = 1, ..., D$. We have

$$\hat{Y}_{d\text{LINPRED}} = \hat{Y}_{d\text{WEIT}} = N_d \tilde{y}_{s_d} + \hat{B}_{d\text{SEP}} (X_d - N_d \tilde{x}_{s_d}) \tag{8.3}$$

where $\hat{B}_{d\text{SEP}} = S_{xyd} / S_{xxd}$. This is a direct estimator, allowing a separate slope and a separate intercept to be fitted in each domain. We can now compare $\hat{Y}_{d\text{LINPRED}}$ and $\hat{Y}_{d\text{WEIT}}$ from two perspectives: (i) the prediction perspective: (ii) the uni-weight perspective.

The prediction perspective: Both (8.1) and (8.3) are linear prediction estimators. They use the same amount of auxiliary information, but differ in the underlying regression model. The model for (8.1) is a regression with a common slope but with separate intercepts for each domain. The model for (8.3) is one in which each domain has a separate slope as well as a separate intercept. It follows that (8.3) has a better fitting model since more parameters are fitted. Consequently, it has a smaller average squared regression residual and usually a smaller variance, compared to (8.1). However, the variance advantage of (8.3) will be highly limited, often one or two percentage points, depending on the population data. For a small domain, (8.1) may in fact be preferred since the separate slope estimate is unstable when based on few data points. Thus, when both (8.1) and (8.3) are available choices (their common auxiliary information is available), the choice between them (which is a choice between two regression models) is not one of crucial importance.

The uni-weight perspective: Both (8.2) and (8.3) are uni-weight estimators. They require different amounts of auxiliary information. An x-total at the population level suffices for (8.2), but the x-total must be known at the domain level for (8.3). Because the information is much stronger for (8.3), it will usually have considerably smaller variance than (8.2). It would in fact be a mistake to choose (8.2) when the information is available to use (8.3). The amount of auxiliary information is more essential than the choice of model.

9. DISCUSSION AND RECOMMENDATIONS

We have argued that the prediction approach is not always practical in a survey with many domains and variables of interest, because the search for the best fitting model will often require a lot of effort. In the much simpler uni-weight approach, we attempt to construct a unique set of weights that give good efficiency for all domains and variables of interest. We have given some formal evidence (theorem 7.1) that it is in our interest to have the information groups as close as possible to the principal domains of interest. The selection of the x-variables and, above all, the specification of the information groups are crucial factors in obtaining high overall efficiency in the uni-weight approach. In this paper we have not addressed a question of considerable importance, namely, how to make sure that high overall efficiency is realized, given the multi-purpose use of the uni-weight system.

ACKNOWLEDGEMENTS

For the formulation of the ideas in this paper, we have benefited from discussions with M. A. Hidiroglou and A. Singh of Statistics Canada, R. Lehtonen of Statistics Finland and S. Lundström of Statistics Sweden. We also acknowledge the helpful comments and suggestions from the Associate Editor and two referees.

APPENDIX A

Proof of Theorem 7.1. For $j = 1, 2$, denote by V_j the Taylor variance of $\hat{Y}_{d\text{WEIT}j} = \sum_x a_k g_{kj} y_{dk}$ where $g_{kj} = 1 + (\boldsymbol{X}_{0j} - \hat{\boldsymbol{X}}_{0j\pi})' \boldsymbol{T}_{s_j}^{-1} \boldsymbol{x}_{0k} / c_k$ is based on the information group U_j, with $\hat{\boldsymbol{X}}_{0j\pi} = \sum_{s_j} a_k \boldsymbol{x}_{0k}$ and $\boldsymbol{T}_{s_j} = \sum_{s_j} a_k \boldsymbol{x}_{0k} \boldsymbol{x}_{0k}' / c_k$ where $s_j = s \cap U_j$. Then we have $V_j = \sum\sum_{(k,l) \in U_j} (a_k a_l / a_{kl} - 1) E_{dkj} E_{dlj}$ where E_{dkj} is a regression residual explained below, and $a_{kl} = 1/\pi_{kl}$ where π_{kl} denotes the probability that units k and l are both included in the sample, and $a_{kk} = a_k = 1/\pi_k$. In general, V_j is a quadratic form in the E_{dkj}, but the expression simplifies for Poisson sampling, where $a_{kl} = a_k a_l$ for all $k \neq l$. Then with only squared terms remaining, we have $V_j = \sum_{U_j} E_{dkj}^2 / Q_k$ where $Q_k = (a_k - 1)^{-1}$. Now since $c_k = Q_k$, we get $E_{dkj} = y_{dk} - \boldsymbol{x}_{0k}' \boldsymbol{B}_{(d)j}$ for $k \in U_j$, where $\boldsymbol{B}_{(d)j} = (\sum_{U_j} \boldsymbol{x}_{0k} \boldsymbol{x}_{0k}' / Q_k)^{-1} \sum_{U_j} \boldsymbol{x}_{0k} y_{dk} / Q_k$. Let $D_{dk} = E_{dk2} - E_{dk1} = \boldsymbol{x}_{0k}' (\boldsymbol{B}_{(d)1} - \boldsymbol{B}_{(d)2})$. Then

$$V_2 - V_1 = \sum_{U_2} E_{dk2}^2 / Q_k - \sum_{U_1} E_{dk1}^2 / Q_k$$

$$= \sum_{U_1} \{(E_{dk1} + D_{dk})^2 - E_{dk1}^2\} / Q_k + \sum_{U_2 - U_1} E_{dk2}^2 / Q_k.$$

It follows from the normal equations that $\sum_{U_1} D_{dk} E_{dk1} / Q_k = (\boldsymbol{B}_{(d)1} - \boldsymbol{B}_{(d)2})' \sum_{U_1} \boldsymbol{x}_{0k} E_{dk1} / Q_k = 0$, so that $V_2 - V_1 = \sum_{U_1} D_{dk}^2 / Q_k + \sum_{U_2 - U_1} E_{dk2}^2 / Q_k$. Because both terms on the right hand side are non-negative, we conclude that $V_2 \geq V_1$.

REFERENCES

ESTEVAO, V.M. (1994). Calculation of *g*-weights Under Calibration and Bound Constraints. Report, Statistics Canada.

ESTEVAO, V.M., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 2, 181-204.

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY (1993). Indirect estimators in Federal programs. Washington, D.C.: Office of Management and Budget, Statistical Policy Office.

FIRTH, D., and BENNETT, K.E. (1997). Robust models in probability sampling. To appear, *Journal of the Royal Statistical Society*, Series B.

FULLER, W. A. (1996). *Introduction to Statistical Time Series*. New York: Wiley.

HIDIROGLOU, M.A. (1991). Structure of the Generalized Estimation System. Report, Statistics Canada.

LEHTONEN, R., and VEIJANEN, A. (1998). Logistic generalized regression estimators. *Survey Methodology Journal*, 24, 51-55.

MONTANARI, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

MURTHY, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

RAJ, D. (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association*, 51, 269-284.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SCHAIBLE, W.L. (1992). Use of small area estimators in U.S. Federal programs. *Proceedings International Scientific Conference on Small Area Statistics and Survey Design*, 1, 95-114. Warsaw: Central Statistical Office of Poland.

SINGH, M.P., GAMBINO, J., and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology Journal*, 20, 3-14.

SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology Journal*, 22, 107-115.

STUKEL, D.M., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: A comparison of Jackknifing versus Taylor linearization. *Survey Methodology Journal*, 22, 117-125.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1999. An asterisk indicates that the person served more than once.

J. Albert, *Bowling Green State University*
C. Alexander, *U.S. Bureau of the Census*
J. Armstrong, *Elections Canada*
B. Bailar, *National Opinion Research Center*
W. Bell, *U.S. Bureau of the Census*
* D.R. Bellhouse, *University of Western Ontario*
Y. Berger, *University of Southampton*
P. Biemer, *Research Triangle Institute*
* D.A. Binder, *Statistics Canada*
* G.J. Brackstone, *Statistics Canada*
K. Brewer, *Australian National University*
J.M. Brick, *Westat, Inc.*
P. J. Cantwell, *U.S. Bureau of the Census*
R.G. Carter, *Statistics Canada*
R. Chambers, *University of Southampton*
G. Chen, *University of Regina*
M. Cohen, *National Center for Education Statistics*
A. Demnati, *Statistics Canada*
* J.-C. Deville, *Institut national de la statistique et des études économiques*
* P. Dick, *Statistics Canada*
A. Dorfman, *U.S. Bureau of Labor Statistics*
J.D. Drew, *Statistics Canada*
* J.L. Eltinge, *Texas A&M University*
V. Estevao, *Statistics Canada*
* W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
J.-F. Gosselin, *Statistics Canada*
* B. Graubard, *U.S. National Cancer Institute*
R.M. Groves, *University of Maryland*
S. Haslett, *Massey University*
* M.A. Hidiroglou, *Statistics Canada*
D. Holt, *Office for National Statistics, U.K.*
B. Hulliger, *Swiss Federal Statistical Office*
J.-S. Hwang, *Academia Sinica*
D. Judkins, *Westat, Inc.*
G. Kalton, *Westat, Inc.*
* P.S. Kott, *National Agricultural Statistics Service*
* M. Kovačević, *Statistics Canada*
J. Kovar, *Statistics Canada*
M. Kramer, *U.S. Bureau of the Census*
R. Lachapelle, *Statistics Canada*
P. Lahiri, *University of Nebraska - Lincoln*
* M. Latouche, *Statistics Canada*
P. Lavallée, *Statistics Canada*
S. Linacre, *Australian Bureau of Statistics*
* S. Lohr, *Arizona State University*
D. Malec, *U.S. Bureau of the Census*
* H. Mantel, *Statistics Canada*
F. Mayda, *Statistics Canada*
P. Merkouris, *Statistics Canada*
J. Moloney, *Statistics Canada*
G.E. Montanari, *University of Pergia, Italy*
B. Nandram, *Worcester Polytechnic Institute*
G. Nathan, *Central Bureau of Statistics, Israel*
A.R. Olsen, *U.S. Environmental Protection Agency*
L. Peng, *Ernst & Young LLP*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
* B. Quenneville, *Statistics Canada*
* E. Rancourt, *Statistics Canada*
* J.N.K. Rao, *Carleton University*
* L.-P. Rivest, *Université Laval*
G. Roberts, *Statistics Canada*
G. Robinson, *U.S. Bureau of the Census*
B. Rosén, *Statistics Sweden*
D. Royce, *Statistics Canada*
D. Rubin, *University of North Carolina - Chapel Hill*
I. Sande, *Bell Communications Research, U.S.A.*
G. Sande, *Sande & Assoc.*
C.-E. Särndal, *Université de Montréal*
O. Sautory, *Institut national de la statistique et des études économiques*
N. Schenker, *University of California - Los Angeles*
F.J. Scheuren, *The Urban Institute*
A. Scott, *University of Auckland*
J. Sedransk, *Case Western Reserve University*
J. Shao, *University of Wisconsin - Madison*
P.L.D.N. Silva, *Instituto Brasileiro de Geografia e Estatistica*
A.C. Singh, *Westat, Inc.*
* M.P. Singh, *Statistics Canada*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
L. Standish, *Statistics Canada*
* D. Stukel, *Statistics Canada*
M. Svershkov, *Hebrew University*
J.-L. Tambay, *Statistics Canada*
* A. Théberge, *Statistics Canada*
* Y. Tillé, *Ecole Nationale de la Statistique et de l'Analyse de l'Information*
C. Tucker, *U.S. Bureau of Labor Statistics*
* R. Valliant, *Westat, Inc.*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
S. Wang, *Texas A&M University*
K.M. Wolter, *National Opinion Research Center*
Y. You, *Statistics Canada*
* E. Zanutto, *University of Pennsylvania*
A. Zaslavsky, *Harvard University*

Past assistant and associate editors:

B. Afonja
J. Armstrong
K.G. Basavarajappa
L. Biggeri
M. Cairns
B.N. Chinnappa
M.J. Colledge
E. B. Dagum
J. Denis
J.-J. Droesbeke
R.E. Fay
J. Gambino
J.F. Gentleman
P.D. Ghangurde
J.H. Gough
M. Gonzalez §
T.M. Jeays
N. Laniel
M. Latouche
P. Lavallée
M. Lawes §
H. Lee
L. Mach
A. Mason
M.N. Murthy
W.M. Podehl
D.B. Rubin
C.-E. Särndal
W.L. Schaible
C.M. Suchindran
J.-L. Tambay
A. Théberge
P.F. Timmons
V. Tremblay

§ Deceased

Acknowledgements are also due to those who assisted during the production of the 1999 issues: J. Beauseigle (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Corbeil, J. Houle, C. Larabie, D. Lemire and G. Ray of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board

Contents
Volume 15, Number 1, 1999

Volume 15, Number 2, 1999

Contents

Volume 15, Number 3, 1999

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS TABLE DES MATIÈRES

Volume 27, No. 3, September/septembre 1999

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.